ETHICS, GOVERNANCE, AND BIAS MITIGATION IN AI-ENABLED IOT APPLICATIONS: A FOCUS ON TRANSPARENCY AND FAIRNESS

NADINE Y. FARES

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF ARTS

GRADUATE PROGRAM IN INFORMATION SYSTEMS AND TECHNOLOGY

YORK UNIVERSITY

TORONTO, ONTARIO

APRIL 2024

Abstract

In the development of smart cities, the integration of Internet of Things (IoT) and Artificial Intelligence (AI) technologies promises to address urban challenges yet raises significant security and ethical concerns. This study emphasizes the need for robust governance frameworks to ensure the ethical deployment of these technologies, focusing on the roles of data privacy, security, and bias mitigation. It highlights the importance of Explainable AI (XAI) techniques, like LIME and SHAP, for enhancing transparency and accountability in AI decision-making processes. Through a detailed examination of existing governance frameworks and the introduction of a bias detection tool, this work aims to embed ethical considerations into the fabric of AI-driven systems in smart cities. The discussion extends to exploring common domains and AI implementations across different industries, offering insights into comprehending and ensuring the responsible use of AI and IoT technology. This lays the groundwork for future research in this developing area.

Acknowledgments

Primarily, a most heartfelt appreciation goes to Dr. Manar Jammal, whose role in this journey transcended that of a supervisor; she was a mentor and a role model. Her unique blend of research acumen, teaching prowess, and exceptional work ethic has made working with her an extraordinary privilege. Dr. Manar's influence extends beyond the confines of this work, inspiring me to strive for excellence in every endeavor and to approach challenges with the same drive and integrity she embodies. There is truly no one else under whose guidance I would have rather embarked on this journey.

I also want to thank my friend Vicky, with whom my friendship began when I was late to one class, and she helped me catch up in the most caring way. From her random check-in calls to our study sessions and outings, she has always been there for me with her positive and uplifting energy.

Additionally, of significant note, I want to express my gratitude to my parents; my mom for being an inspiration of strength, virtue, and independence, and my dad for being my guiding light, always there with unwavering love and support. Their constant trust in my vision has been a cornerstone of my journey, from career changes and venturing into new fields to relocating abroad and beyond.

And finally, my sister Nour, to whom I dedicate this work, thank you for being my rock, my foremost listener, and my lifelong best friend.

Table of Contents

List of Tables

"All tables presented in this document are original works created by the author and have not been taken from external sources unless otherwise noted."

List of Figures

"All figures presented in this document are original works created by the author and have not been taken from external sources unless otherwise noted."

Preface

The research presented in this thesis is the original work of Nadine Fares, and parts of it have been accepted to the peer-reviewed publication listed below:

- Fares, N. Y., Nedeljkovic, D., Jammal, M. (2023) “AI-enabled IoT Applications: Towards a Transparent Governance Framework” Accepted at IEEE 7th Global Conference on Artificial Intelligence and Internet of Things.

- Fares, N. Y., & Jammal, M. (2023). AI-Driven IoT Systems and Corresponding Ethical Issues. In AI and Society (pp. 233-248). Chapman and Hall.

Parts of this research are to be submitted for publication:

- Fares, N. Y., & Jammal, M. “Decoding Justice: The Synergy of Artificial Intelligence and Machine Learning in the Legal Landscape.” Submitted to International Conference on Artificial Intelligence Applications in Law (ICAIAL -24).
- Fares, N. Y., Moore, S., Jammal, M. “Exploring Bias in Recruitment: Utilizing LIME and SHAP for Transparent Machine Learning Predictions”. Submitted to IEEE International Symposium on Technology and Society (ISTAS 2024).
- Fares, N. Y., Moore, S, Jammal, M. “Unveiling Fairness in Credit Decisions: Leveraging LIME and SHAP for Bias Detection in ML Models”. Submitted to 12th IEEE International Conference on Intelligent Systems.

Abbreviations

AI **Artificial Intelligence**

AMA **Artificial Moral Agent**

CSA **Cloud Security Alliance**

DL **Deep Learning**

FN **False Negative**

FP **False Positive**

IoT **Internet of Things**

IMDA **Infocomm Media Development Authority**

ISACA **Information Systems Audit and Control Association**

ISAGO **Implementation and Self-Assessment Guide for Organizations**

LIME **Local Interpretable Model Agnostic Explanations**

ML **Machine Learning**

NIST **National Institute of Standards and Technology**

PDPC **Personal Data Protection Commission**

RFID **Radio Frequency Identification**

SHAP **Shapley Additive exPlanations**

SVC **Support Vector Classifier**

TP **True Positive**

XAI **Explainable AI**

1. Introduction

In the realm of smart cities, the integration of advanced technologies aims to alleviate urban challenges with minimal human intervention. Addressing concerns like population growth, waste management, traffic congestion, resource accessibility, and crime, these initiatives stand as promising solutions. However, amidst the spotlight on these modern technological advancements, the discourse around ensuring adequate security measures are in place for the development of smart cities remains insufficient.

Internet of Things (IoT) serves as a facilitator of smart cities. Data, on the other hand, stands as the cornerstone of smart cities, drawn from diverse sources like surveillance devices, RFID tags, financial records, and sensors, waiting to be connected and transmitted through IoT. Leveraging this data necessitates the implementation of sophisticated surveillance tech and enhancements to traditional devices and legal frameworks, ensuring the collection of precise and consistent information from these smart communities. Consequently, security systems harness advanced technologies to extract valuable insights and bolster security measures.

The advent of Artificial Intelligence (AI) has significantly transformed various smart city development strategies, offering the potential to augment predictive accuracy and optimize resource allocation. However, amidst these advancements, it's crucial to acknowledge and mitigate associated risks, such as biases, that could potentially infiltrate these AI systems. The

escalating integration of AI and IoT in smart cities demands robust governance frameworks and cautious innovative steps to uphold transparency, fairness, and ethical use.

1.1) Thesis Objectives

An integration of AI in IoT necessitates a concerted effort to detect, mitigate, and rectify biases at every stage of AI development and deployment within smart city initiatives. Techniques like Explainable AI (XAI), notably Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), along with other measures, assume critical roles in addressing these challenges. They provide invaluable insights and clarity into the decision-making processes of AI, promoting accountability and ethical considerations. Our objective extends beyond mere evaluation of these approaches; the goal is to help develop methodologies capable of automatically assessing AI applications for compliance with ethical standards and regulations. This proactive approach aims to embed accountability and ethical scrutiny within the very fabric of AI-driven systems in smart cities, ensuring their responsible and equitable deployment.

Throughout this work, we tend to navigate and answer the following questions:

- How do AI-enabled IoT applications intersect with ethical and legal risks, such as bias and lack of clarity?
- What governance and supervision mechanisms are necessary to mitigate these risks?
- What are the limitations of current frameworks governing AI and IoT applications?

- How can Explainable AI (XAI) strategies enhance oversight and accountability?
- How effective are XAI techniques, specifically LIME and SHAP, in interpreting machine learning models' decision-making processes to identify and quantify biases?
- Can the introduction of a risk analysis tool automate the bias-detection process in machine learning predictions, thereby ensuring greater transparency, fairness, and accuracy?
- What legal considerations are paramount for the safe and ethical deployment of AI technologies?
- How can insights from the legal domain inform best practices across various industries?
- What are the implications for future research and development in the field of AI and IoT, particularly in enhancing ethical and legal compliance?

1.2) Thesis Outline

This work is divided into multiple chapters. Chapter 2 commences with a general exploration of AI-enabled IoT applications moving to how they associate with ethical and legal risks like bias and unclarity, and their need for proper governance and supervision. Chapter 3 discusses the traditional frameworks currently governing AI and IoT along with XAI strategies that can help further oversee AI-IoT-enabled applications. In Chapter 4, XAI techniques, such as (LIME) and (SHAP), are further inspected along with other investigative calculations and applied to interpret ML models' decision-making processes which will unveil the features driving predictions,

enabling the identification and quantification of biases. Moreover, we introduce a risk analysis tool that will automate the bias-detection process and ensure a more transparent, fair, and accurate ML prediction. Chapter 5 then delves specifically into the domain of law and its applications concerning the safe and ethical deployment of AI. This provides insights that can inform other domains on best practices for implementing AI in their respective practices or services. Chapters 3, 4, and 5 also encompass an in-depth analysis of prior works or contributions in the specified areas. Finally, the conclusion highlights the work's significance and outlines potential future research directions.

2) AI and IoT Background: Opportunities and Risks

2.1) Introduction

Artificial Intelligence (AI) is already gaining major attention among societies. The enduring development of automation offered a wide range of technological services and applications within other advanced systems, one of which is the Internet of Things (IoT). While AI enables devices to analyze and learn from experiences, IoT uses devices to connect people and objects through the internet. With efforts to render IoT systems AI-enabled and automated, attention should be given to the potential ethical risks and concerns alongside the designed enhancing features and the benefits predicted. The synergy of AI and IoT can result in advanced and smart applications within smart cities, driverless vehicles, and robotics and can initiate new job markets and demands. However, the converged system can also threaten general ethical norms and human rights through algorithm bias, cybercrimes, unemployment, liability concerns, and autonomy questioning. This chapter will explain both systems, AI and IoT, in terms of definitions, history, and methods and elaborate on the advantages, disadvantages, and corresponding recommendations of their union to offer a better understanding of the advanced AI-enabled IoT system and reach a clear resolution regarding the matter.

2.2) IoT and AI: Background

In this section, we will provide a brief background of each of AI and IoT, along with an explanation of relevant definitions.

2.2.1) What is Artificial Intelligence (AI)?

AI uses algorithms within machines and focuses on giving them cognitive abilities such as reasoning, problem-solving, decision-making, and recognition [4]. Ever since its creation, AI has evolved over the years while undergoing cycles of both praise and criticism. The field came into existence in 1952 with advances and automation in neuroscience, attracting the focus of both the scientific industry and the public community [5]. Then in 1956, the Dartmouth Conference positively embarked on the era of AI. The AI journey started with manufacturing computers that could solve various mathematical problems. Since the technology was burdened with limitations at the time, through public doubts in efficiency and high costs, the public questioned whether the hype over AI was exaggerated, despite the promising features of this field. As a result, several agencies halted their funding services to AI, leading to a wave of pessimism against technology [6]. The field faced its public ups and downs but kept progressing as computer power and resources further developed. Finally, the rise of the Big Data revolution, which involved big volumes of data of wide varieties and high velocities, made data accessible enough to advance AI and initiate its application through various sub-domains such as “machine learning” and “deep learning”. According to literature, while machine learning focuses on developing computers to work with fewer human interventions, deep learning is a subset of machine learning that focuses on giving computers the ability to “think” using structures modelled on the human brain [6]. The availability of data and resources allowed for the actual implementation of theoretical AI/ML/DL in real-life applications such as Google filter spams and image detection and led to the debatable

idea of technological singularity; predictions of the emergence of AI that surpasses human intelligence with that of machines [5].

AI gained wide attention of academic and popular literature and research. Actual and practical applications of AI have been generated and used in the fields of manufacturing, cybersecurity, health, cars and transportation, and many more by worldwide famous companies like the IBM's Watson, Tesla's autopilot, and Google applications.[4]

2.2.2) Weak vs Strong AI

In terms of capacity, AI is categorized by computer scientists as either weak or strong AI. By definition, weak AI, also known as narrow AI, is the most common among current intelligent systems and is trained to perform and repeat observed behaviors and assigned tasks [5]. Weak AI relies on humans to define the limitations of the learning processes and to deliver the relevant training data [7]. Despite lacking the ability to generalize, weak AI is evidently efficient in machine learning, data mining, language processing, and pattern recognition. Systems that are supported by weak AI include self-driving cars, recommender systems, virtual assistants (Siri), spam filters, and industrial robots [5].

On the other hand, researchers describe strong AI to be complemented by a sense of human-like consciousness and ability to reason and think [7]. According to existing literature, in addition to being able to integrate information like the weak AI systems, strong AI can autonomously reprogram itself and modify its functioning to perform general tasks that require intelligence, sentience, and "self-awareness". According to IBM cloud education, strong AI can teach itself to solve new problems by developing a human-like consciousness instead of imitating it [7].

Despite the difference in autonomy and control between weak and strong AI, both systems represent the common feature of independency from full human control. Consequently, both systems carry common challenges accompanying their AI-enabled applications and are of the same importance when discussing ethical concerns and legal matter.

2.2.3) What is Internet of Things (IoT)?

IoT systems have not been around for a long time, but the vision behind their technologies had already been under development many years ago. In 1969, the Internet, the key component of IoT, developed as Advanced Research Project Agency Network (ARPANET) and was mainly used for research and academic purposes. In 1973 and 1974, other essential technologies for IoT, respectively Radio-Frequency Identification RFID and embedded computer systems, were established and implemented for various uses [8]. Despite the proliferation of these technologies in businesses and markets, IoT was not recognized as an official term only until 1999 by Kevin Ashton and was not applied as an individual system only until the 2000s with internet connectivity becoming an essential norm for most applications and products [9].

There exist several definitions of the term IoT. According to Internet Architecture Board (IAB), IoT is a set of a large number of embedded devices termed as “smart objects”, which communicate with each other without human intervention to provide communication services [8]. The Internet Engineering Task Force (IETF) refers to IoT ‘Smart Object’ too, but of limited power, processing ability, and memory. Moreover, IEEE Communications Magazine relates IoT to Cloud Services as a framework that aims to offer several services, using Machine-to-Machine M2M communications, to eradicate the gap between the physical and the virtual world. Atzori et al. define Internet of Things as three key concepts that intersect with each other to develop the full potential of IoT: internet-oriented Middleware, object-oriented Sensors, and semantic-

oriented Knowledge [8]. Oxford Dictionaries defines IoT very accurately as “the interconnection via the Internet of computing devices embedded in everyday objects, enabling them to send and receive data”. Consequently, a general and inclusive definition of IoT would be: a smart environment that utilizes the Internet, information, and communications technology to offer services to several domains like administration, education, healthcare, and transportation. With a framework that includes global and cloud computing for sharing and accessing data, data analytics for storing and analyzing data, and knowledge visualization for representing data, IoT has changed and will keep changing the entire communication process [8].

Since IoT systems are growing at high speed, it is important to know the several communication models in which IoT devices connect and intersect in. According to a document prepared by Internet Architecture Board (IAB) for networking of smart objects, there are four major communication modules behind IoT systems [10]:

a) Device-to-Device Communication Model: Communication is directly between two or more IoT devices through Internet or any other network, with no need for an intermediary application server. Some examples include home automation systems lie Amazon Alexa and Google Nest Mini.

b) Device-to-Cloud Communication Model: Communication is between the internet cloud services and IoT device. It is done through traditional wired Ethernet or Wi-Fi connections to connect the device to the IP network, which consequently connects it to the cloud service. Some examples include Samsung Smart TV and Nest Labs learning Thermostat.

c) Device-to-Gateway Communication Model: A gateway device, also known as application-layer-device, operates through a software and connects the IoT device to the internet cloud services for the data exchange. Some examples include Fitness applications in a smart phone, which acts as an application layer gateway - it is connected to the fitness tracker device which cannot directly connect to the clouds service.

d) Back-End-Data-Sharing Communication Model: Communication is driven by users sharing/exporting the IoT device sensor data stored in the cloud database and joining it with third party data. This model overcomes the limitations of device-to-cloud model and allows the analysis of data collected from several IoT devices. Moreover, back-end data-sharing model was derived from single device-to-cloud communication model [11].

2.3) Advantages of AI-enabled IoT systems

Recent advances in connectivity and technology have led to the emergence of AI and IoT systems in many industries [12]. While IoT is responsible for collecting data and visualizing it, AI plays a significant role in the analysis of expanding amounts of data, learning from them, and taking action accordingly. Consequently, the users, sensors, and networks of IoT produce huge amounts of data from which professionals can acquire knowledge and develop applications using AI-enabled methods [13].

However, as technological dependency increases with time, the number of companies utilizing IoT-data are becoming countless, which complicates the process of the proper collection, analysis, and acting upon the data. According to specialists, AI is the most suitable tool to understand the huge amounts of data and patterns and produce informed decisions and efficient

results in a short timeframe [2]. The synergy of AI and IoT can reshape the methodologies of businesses, industries, and economies. Here are some results of use cases where AI has been used along with IoT applications.

2.3.1) Transforming Job Markets and Creating New Products & Services

In parallel to innovation, various relevant occupations will emerge in which humans will conquer automation due to their judgement, knowledge, and critical thinking abilities [14]. Careers in the domains of science, computer science, health, communications, and education will become those in highest demand - these include data scientists, data administrators, tech-educators, and robotic specialists. Also, NLP (Natural Language Processing), a subfield of AI, allows the communication of people with devices. Undeniably, by installing this feature in IoT systems, the business can promptly process and analyze larger amounts of data, enhance existing products, and create new products and services [2]. For instance, PwC designed a structure to take in drone imagery of construction sites, allowing construction progress to be compared with the planned quality and quality [3].

2.3.2) Better Risks Management

The collaboration of AI and IoT helps businesses to better understand and predict a wide range of risks and create a fast and automated response to fix and moderate. Given the ability of AI to analyze large amounts of data, insights are more possible to be achieved. This consequently allows companies to properly manage cases of financial loss, cyber threats, and employee safety

and security [2]. For example, Fujitsu installs protective measures for their workers by having AI analyze data taken only from connected wearable devices [2].

2.3.3) AI-enabled Robots

AI robotic systems do not require pauses or breaks as they are developed and programmed to work for long hours and achieve better results in a short timeframe. Moreover, robots are not possessive of human-like emotional side and that prevents room for timely distractions. Robots and machines depend on logical thinking, can work on more complex and laborious tasks, and give direct responses to arising matters when needed [15]. These characteristics, however, drive several ethical concerns that will be discussed later. Examples of AI-enabled robots include IBM Watson, DaVinci Surgical System, and nanorobots [5].

2.3.4) Personalized experiences

The junction of AI and IoT leads to systems working with customers/users at a personal level. Interactions between the user and the system produce personalized results according to the user's specific desires and details, such as health monitoring by wearable devices, advice by recommender systems, peace of mind with smart household products, and convenience with virtual assistants [16]. These applications include fitness trackers, Google Assistant, Apple's SIRI, Amazon's Alexa, etc [17].

2.3.5) Reduction of error

AI may reduce the chances of error and is developed to reach accuracy and precision in a smaller number of attempts in comparison to traditional approaches [18]. For example, several weather forecasting groups use AI to reduce the disadvantage of human error. Also, robotic surgeons, such as the DaVinci System are designed to be associated with movement precision during medical procedures [5].

2.3.6) Telehealth

According to Rushabh Shah et al, the major complications associated with medicine are disease diagnosis and treatment, health monitoring, data collection and analysis, and accuracy [12]. One of the IoT applications is telehealth - tracking and monitoring of patients' health conditions through wearable devices, biochips, and remote diagnosis. With the addition of AI technologies, such as neural networks, healthcare systems can provide physicians with specific patient information and historical facts, allow for proper analysis of the data, and rationalize the diagnosis and relevant treatment in an accurate and risk-free manner [19].

2.3.7) Smart Cities and Smart Governments

IoT-sourced data include that of smart cities and governments. AI-enabled IoT applications can be applied to improve the development of smart cities, expand efficiency of smart governments, and provide life-

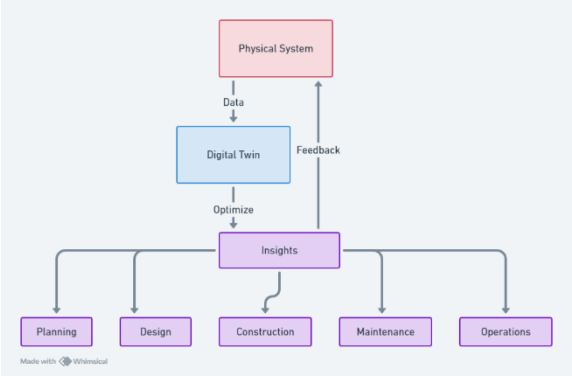

Figure 1 - Making Data Valuable

enhancing services for citizens [13]. Some of the major areas of a smart city are the smart manufacturing, smart transport system, smart energy, and smart homes. In smart manufacturing, the convergence of the physical and virtual worlds led to many developments, one of which is digital twin. Digital twin is a concept that involves a virtual model of a physical asset for predictive purposes [20]. It operates by real-time sensory data and can estimate the potential issues of the respective physical assets (Figure 1). In IoT world, AI will enhance the efficiency of digital twins through enhanced data analysis features that better understand states, add value, improves operations, and respond to changes.

2.3.8) Elimination of costly unplanned downtime

In some sectors like industrial manufacturing, equipment breakdown can cause costly unexpected stoppage. The predictive preservation offered by AI enabled IoT allows businesses to predict the failure or complication of equipment beforehand and plan timely maintenance procedures. Intelligent IoT systems can utilize AI to predict the conditions of operations and detect the parameters that need to be modified to ensure ideal results. Hence, efficiency can be enhanced in tasks that possess time-consuming maintenance processes [2].

Deloitte, a multinational audit and consulting firm, finds that upon integrating AI and IoT in their system, they achieved "'20% - 50%' and '5%-10%' reductions in their time invested in maintenance planning and costs respectively and '10% - 20%' increase in equipment availability and uptime" [2].

2.3.9) Self-driving cars

Self-driving cars are a famous example of IoT applications. A self-driving car requires an immense quantity of data collection and processing. These data involve information about the road, traffic, and navigation. Artificial Intelligence can play a role in achieving efficient data analysis and reaching better driving and route decisions. According to Lawrence Burns, self-driving vehicles are significantly lighter and can reduce out-of-the-pocket and time costs of conventional automobile travel by 80%. Also, his research at Columbia University concluded that automobility is the sustainable solution to the enormous waste inherent in human-driven, combustion-powered, manual automobiles [21].

2.4) Ethical Challenges of Enabling AI in IoT Systems:

Alongside the researchers who encourage the integration of AI into IoT applications due to its promising impacts, others are hesitant about the manner due to the ethical issues and conflicts that arise in parallel to innovative projects. The synergy of AI and IoT raises several ethical concepts that need to be addressed:

2.4.1) Robo-ethics vs AI singularity

With manufacturing robots joining several workforces, it is crucial to discuss the ethical standards of robotics, also noted as “Roboethics”, and how they should be implemented in AI-powered IoT robotic systems. Scientist and futurist Isaac Asimov developed three initial laws

(1950) followed by a zeroth law (1985) for the field of Robotics [22]. From lower to higher-order, the laws were:

- A robot should not harm humanity
- A robot may not injure a human being or, through inaction, allow a human being to encounter to harm.
- A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law.
- A robot must protect its own existence if such protection does not conflict with the First or Second Law.

In 2002, ethicists and engineers tackled this field again to widen the scope of the previously set guidelines, develop more specific ones, and clearly distinguish between similar concepts. For example, while Roboethics focused on the rules to be considered in the human manufacturing of robots, “machine ethics”, in the contrast, focused on the morality of machines and considered artificially intelligent machines as “Artificial Moral Agents” (AMAs) [22].

In 2011, the Engineering and Physical Sciences Research Council (EPRSC) and the Arts and Humanities Research Council (AHRC) of Great Britain issued a set of standards for robot designers to follow to guarantee common and shared values to produce robots. These standards emphasize that robots, as multifunctional tools, should prioritize safety, security, and ethical use, with clear accountability and adherence to laws and rights, avoiding harm or deception towards humans [22].

It is important to notice that despite the fantasized unlimited power associated with AI, and consequently with its subfields such as robotics, the aforementioned updated robot production regulations were generated with the preconception that robots are eventually accountable to pre-set human laws. This, however, clashes with AI singularity; the notion that refers to a world where everything is performed by ultra-intelligent machines and autonomous systems that can overpass human intellectual activities and rules [23]. While some urge to draw attention to the seriousness of the situation, others claim that this technological individuality is still far from close. Yet, there is no harm in early preparation and protection. The implementation of AI robots in IoT systems requires serious debates on the manufacture and designing of the robots and their corresponding abilities and limits. This is because, in robotic developments, ethical and legal complications expand as the autonomy of devices increases [24].

2.4.2) AI bias and discrimination

AI implementation does not only affect the work aspects of humans, but also their lifestyles. Al algorithms and training information and data may repeat, reinforce, or augment harmful biases since these data are also generated by humans. Consequently, these biases drive AI systems to produce unfair outcomes and discrimination against minority groups. Bias in AI algorithms is caused by two reasons: 1) Developers would insert biased information in AI systems accidentally and without even noticing; 2) Historical data that will train AI algorithms may not be wide enough to represent the whole population in a fair manner [25]. For example, in search engines and translation algorithms, outputs are largely perceived as objective, yet frequently encode language in gendered ways. Also, bias has also been evident in the algorithmic advertisement, with opportunities for higher-paying jobs and jobs within the field of science and technology

advertised to men more often than to women. Likewise, prediction algorithms used to manage the health data of millions of patients give better care to white patients in comparison to black patients [26]. The risk of Bias will be further explored and focused on in the subsequent chapters.

2.4.3) Cybercrimes, data breach, and lethal weapons

While AI can be used by cybersecurity professionals to enhance data protection and minimize breach attacks in IoT systems, cybercriminals can abuse AI abilities for malicious and adversary purposes. This could lead to developing machine learning models that misinterpret inputs into the system and behave in a way that would make the breach or cyberattack easier and more favorable to the attacker [27].

As mentioned before, AI offers the advantage of personalized experiences within IoT facilities in customer services and more. However, with AI systems being able to process personal data, individuals' rights to privacy and freedom are at risk. This is because in AI systems, data collection abilities driven by low storage costs lead to data persisting longer than the human subjects that created it. Moreover, the data collected can be abused or used for purposes beyond their primary expected ones without the data owners' consent. Also, data may be collected from people who are not the original targets of the data collection sequence leading to a breach in data integrity [28].

AI can be integrated into systems for criminal and harmful purposes. An example is incorporating AI in the military and developing lethal automated weapons LAWs that could independently identify targets based on programmed descriptions and constraints. There lie

ethical concerns with this practice, especially with the risk of bias and discrimination associated with the AI systems. [25].

2.4.4) Unemployment

While automation with AI is expected to relieve humans from complex and unnecessary fatiguing IoT tasks, working individuals fear that their jobs will be eliminated by AI in the future. According to Mckinsey estimates, smart robots and intelligent machines could replace as much as 30% of the current human labor globally by 2030. Depending upon numerous implementation scenarios, automation will displace between 400 and 800 million jobs, and about 375 million working individuals will be required to switch job categories entirely [25].

2.4.5) Disrupting autonomy and human relationships

Self-rule, also known as self-governance or self-determination, is the major factor of autonomy. Smart objects, unlike traditional manual objects, acquire their own unique capacities and their own kinds of experiences in interaction with the consumer and each other [29].

Moreover, AI advancements are currently being used to dynamically personalize individual's choice scenarios, paternalistically nudge, and manipulate behaviour in unmatched manners [30]. Expanding deference to algorithmic patterns in many decision-making processes creates doubts regarding the authenticity of individuals' decisions, and consequently their capacity for self-determination and awareness. This is even more evident when algorithmic decision-making patterns are dependent on group-level data and past information.

Moreover, Artificial bots can virtually channel unlimited attention and kindness into building relationships, unlike humans, who expend these resources in a limited manner. This reward/fulfillment feeling will result in the reduction of human communication and an increase in human-bot interaction [31].

2.4.6) Duty of care and liability

The self-learning abilities of AI systems can help them come up with entirely new decisions and data on their own. While a key part of negligent liability is foreseeability, autonomous systems are likely to display unforeseen behaviors, and this would present challenging dilemmas in determining who is liable [32]. This could then lead to unforeseeable harms or drawbacks that were not predicted by the developers. This creates ethically controversial matters; while some claim that no matter how automated the system is, human developers and operators hold a duty of care towards the decisions reached by the AI systems they programmed, others suggest that autonomous machines should be granted legal personalities and thus liability in cases of negligence and harm [33]. Generally now, humans are to be responsible for mishaps. However, this predetermined human liability is harsh, as there would exist cases where harm occurs even if all the groups involved act in an ethical and legal manner and with intentions to benefit rather than to harm. Yet, it would still be unfair for the impaired victim/company to not be compensated for the harm they would have suffered [34].

2.4.7) Deteriorating social equity and disparities in wealth and power

Deploying AI in IoT systems is costly and relatively complicated, which makes it an applicable practice for limited countries and advanced communities. If these autonomous systems were to be sent by advanced countries and deployed in disadvantaged or developing countries, the process of implementation would lead to several complications and would eventually fail, as these systems would not only be expected to perform new and difficult tasks but would also have to conform to undeveloped and irrelevant regulations and legislations [24].

Also, as mentioned before, AI-powered systems are expected to correlate with unemployment, so traditional human workforce is predicted to suffer from a cut-down, making revenues go to fewer people. Consequently, rich individuals who have partnerships in AI-based IoT-driven companies will receive wealth and power. This makes advancing IoT with AI associated with further exacerbating the existing gap between privileged and underprivileged communities in the world despite that increasing connection and world communication is one of IoT's major goals [31].

2.5) Open Discussion

The integration of artificial intelligence in IoT systems introduces a wide range of ethical and regulatory issues. Different perspectives exist regarding the matter, which proves that there is an urgent need to reach a united position. However, as the world is getting deeply digitized, ignoring the trend of innovation is not preferred. Steps should be taken to overcome challenges and technology should be encouraged, rather than limited, especially when it seems to hold benefits and promising outcomes.

According to Travis Hollman, CEO of Hollman, the world's largest manufacturer of wood and laminate lockers, robots are best to fill in for humans in repetitive tasks, associated with high-volume production, and are often considered dangerous or boring [35]. This way, AI-IoT manufacturing robots would save people from injuries and relieve stress on the company or the institution.

In parallel to the disputes over the rights and wrongs of employing smart machines instead of human workers, research makes it apparent that companies that fail to embrace the AI opportunity will have to face significant disadvantages that no team can afford. And with the acceleration of AI and IoT implementation, the chances of finding competitive alternatives and weighing options are quickly diminishing [3].

IoT and AI, together, can enable the development of valuable services for citizens, businesses, and public agencies, in multiple domains, such as transportation, energy, healthcare, education, and public safety [13]. Nevertheless, innovative legislations and clear standards and duties, for all parties involved- workers, manufacturers, software designers, and administrative team - should be clearly set out to eliminate the misuses of AI and to help ease the mentioned ethical, legal, and regulatory challenges and other relative issues that would arise.

Moreover, while it is arguable that the autonomy advantage of AI may not be strictly necessary, as it limits communication and human interactions, the autonomous feature would still be advantageous and provide an extra level of distancing and protection to individuals in specific situations [24]. For instance, the recent and ongoing Coronavirus pandemic has majorly highlighted the need for smart systems and autonomous robotic systems to minimize the dangers of exposure to the highly contagious virus for the working staff and other individuals [36].

Therefore, after weighing the pros and cons of integrating AI features into IoT systems, it is obvious that the advantages of incorporating AI in IoT outweigh the disadvantages, and thus, embracing this Hi-Tech collaboration is better than ignoring it. However, for protective measures, it is best to support the following position: AI-powered IoT systems should be supported only if they were to be viewed as a regulatory moderated source of assistance and strength to humans and their businesses, rather than a threatening factor or unethical replacement. Pierre Dupont, an engineer in Boston Children's Hospital's robotic medical research team, emphasizes the fact that autonomous systems should be designed to support human workers and not outshine them [37]. This perspective can also be implemented outside the medical scope; in most businesses and firms that are willing to incorporate an AI-enabled IoT system to manage their industrial work and increase their productivity.

2.6) General Considerations

On one hand, businesses and people should not be discouraged from working with AI and developing innovative technology, especially when current AI devices have been used in many cases successfully [38]. On the other hand, users utilizing the IoT devices/services must undertake all reasonable measures of minimizing the risks and concerns associated with this technology. Therefore, it is proposed to keep the following recommendations in consideration:

- If we plan to encourage the progress of AI and further develop it in most systems, it is important to first understand what practicality and science fiction are [5]. Hence, there should be a logical and smooth interaction between visionary research ideas and the actual application of specific projects. Moreover, it is required to preclude the unrealistic

enthusiasm and the unjustified fears of AI from hindering its progress. Instead, these overhyped feelings and visions should be used to inspire the development of a systematic outline on which the future of AI will prosper. If a sustained and a responsible investment is set, AI is expected to transform the future of society as a whole – people's life, their living environment, and the economy. AI and Robotics, if handled adequately, could amplify and augment human potential and improve productivity [5]. Yet, to properly manage the influence of AI, it is important to understand and draw lessons from both past failures and achievements.

- One of the major complications associated with AI-enabled IoT systems is the issue of the projected ethical and legal challenges they would further impose on the 'current' ethical and legal complications of the advanced technology. After clearly identifying all the impediments, adequate steps should be taken to tackle them. Ethicists and professionals should encourage constant comprised and communication to reach a united and clear ethical framework regarding the application of advanced systems. In addition, since the law is known to usually fall behind against the racing technological advancements, either altering current laws or setting new innovative legislations, or even both, is/are needed to allow the progress of technology and AI [24].
- Since the physical, mental, and digital safety of individuals mark as priority to both the industrial and business systems, scientists and professionals are required to further study the factual aspects of incorporating AI in IoT systems and properly assess the results in comparison to the traditional procedure. Therefore, more experiments and research on the subject are needed before fully transforming systems using AI [37]. This is important to ensure the safety of individuals and clarify the expected standards of both the traditional

and advanced digital systems. Consequently, this will support the achievement of the best decisions for ultimate outcomes and will aid in improving areas of shortage or risks.

- To guarantee proper utilization of AI in IoT systems, clear regulatory measures should be created. Guidelines, codes of practice, conditions, and principles should be followed by manufacturing firms and adopting institutions to ensure that the process of designing and utilizing these advanced systems is not abused for profit purposes at the expense of human rights or even exploited by the staff to escape major duties [39]. Both non-binding and binding measures should be instituted; the non-governmental nature of the former would allow the measures to be trans-national and thus, more flexible and suited to this form of technology, whereas the latter would ensure judicial protection of rights and justice. Moreover, a regulatory framework should be formed to target the hostile and risky environments, those that would seem in favor of automated and advanced systems. This is because in health, for instance, doctors would still have a duty of care and responsibility towards their patients. It should be clarified and agreed that they are willing to accept the idea of relying on an electric robotic system they cannot directly control [24]. The aim is to guide private sector organizations and promote public understanding and trust in technologies by providing a clear explanation of how AI systems work, building adequate data accountability practices, offering transparent communication, and redesigning jobs to fit with the technical advancements [40].
- Technological advancements should stay in compliance with international human rights. Therefore, with the entry of costly and complicated innovative services into various work fields, rulers should ensure that the access of individuals to these services is equal and fair. In a world suffering from disparities across communities, keeping the aspect of

"equal access" to services should be a goal in technological planning. Since every human has an international human right to the ultimate form of health, privacy, and a life of dignity, innovative advancements should bolster these rights, rather than threaten them. It would also be preferable to focus on easing or even abolishing current disparities before introducing advanced hi-tech services that would seem to further exacerbate the situation.

2.7) Summary

The enduring development of AI offered a wide range of technological services in society [22]. Advances in AI have undergone rapid growth over the past decades. They have gained extensive popularity in most fields and have been applied in many practices. AI systems have provided comfort in IoT areas where manual control and tasks were challenging. Moreover, they have provided a benefit of procedural continuity, as they have eased work fatigue and unnecessary exhaustion that could stand in the way of other achievements [41]. However, in parallel to the rise of aspirations and optimism with its successful results, AI-enabled IoT systems have posed major ethical and legal concerns. And with the efforts to render more AI-integrated systems feasible, all involved parties- governments, policymakers, enterprises, innovators, developers, and citizens are challenged to be impacted by the intricate ethical and legal complications associated with this revolution, especially when it comes to privacy and autonomy rights, ensuring responsibilities of individuals and firms, and securing equity and justice [32].

Therefore, companies that are willing to utilize this proactive approach should engage in extensive monitoring and moderating procedures to ensure proper practice and limit ethical risks and misuses. This way, they could make the most of the opportunities offered by AI while

securing the rights and safety of the involved individuals- employees and utilizers.[3] Including AI in IoT systems, just like in all other systems, is a major responsibility and not an on-off switch; once it is installed, it is hard to reverse the impact. To take this responsibility, adopters are urged to have a foundation of knowledge of AI to adequately understand how it may impact not only the adopting system, but the society as a whole.[4]

This chapter sets the stage for the next one, which further explores the suggested recommendations regarding the need for transparency and proper governance in AI. The first step involves assessing existing frameworks, evaluating them, and introducing methods to comprehend AI models for effective governance (known as XAI).

3. AI-enabled IoT Applications: Towards a Transparent Governance Framework

3.1) Introduction

Smart cities, self-driving cars, and smart farming are all examples of Internet of Things (IoT) applications. These applications primarily focus on offering practical services by connecting devices to humans or other devices with minimal to no commanding intervention of humans. It is estimated that by 2025, the number of connected devices will reach 100 billion [42].

Artificial Intelligence (AI) has been distinguished as a revolutionary factor in human and industry services. Since AI aims to emulate human abilities in analysis and decision-making, IoT applications can use AI to further advance their results and maximize their independence from humans. However, as with all technological advancements, legal and ethical concerns increase with the decrease of human control. These concerns relate to bias and data discrimination, cybersecurity issues, unemployment, vague liability, and unclarity. This leads to a focused need for governance, management frameworks, and regulations.

In this chapter, and in the context of architecture and the need for management, we explore some IoT applications, with the available literature and opinions on their administration and the current governance framework models of each AI and IoT. We realize that so far, the lack of a specific governance plan and the need for a detailed and inclusive one can be generally agreed on by most researchers in the domain. From there, we evaluate the available policies and assess model

frameworks for governing AI-driven IoT applications for transparency and better AI-decision making. Our chapter contributions can be summarized as follows:

- Clarify that there is no “one size fits all” methodology to govern advanced technical systems.
- Explain how governance frameworks on IoT and AI systems vary from the initial design phase to the final implementation phase and beyond.
- Explore methods to ensure transparency and trust in AI, which is key to ensure safe practice and fair communication.
- Emphasize the need for maintaining a human-centered approach in AI and IoT systems, regardless of how advanced technology gets.

3.2) AI-based IoT Applications: Background

In this chapter, we present our ideas using an informative and qualitative approach. We explore different studies and research papers on IoT that thoroughly analyze different IoT applications to gain insights into the perspectives of scholars on IoT applications and the integration of AI in them. We then focus on studying existing AI and IoT policies; ones that deal with the design phase of these advanced systems, ones that deal with the implementation phase, and ones that deal with both design and implementation and beyond. Currently, different policies are placed in action to govern IoT and AI systems. They include: Singapore’s Infocom Media Development Authority IMDA IoT Cybersecurity Guide, the National Institute of Standards and Technology NIST, and PDPC The Model AI Governance Framework.

The swift digitization of common traditional services is moving communities towards data-driven activities and decisions [43]. This movement, however, should be secure and safe to attain the best results. Although IoT applications are numerous, we will focus here on four key applications: Smart cities, Self-driving vehicles, and Smart farming. These applications constitute common and popular use in many technologically oriented societies, and for that, they require high governing attention and need for effective ethical and operational regulations.

A. Smart Cities

IoT-sourced data include that of smart cities and governments. Smart cities include major smart services and domains like smart manufacturing, smart transportation systems, smart policing and crime prevention, smart energy, and smart homes [13]. Lopes claims that developing smart and innovative cities can only be done with a smart governance model, a mix of collaborative, open and participatory governance [44]. Joshi et al. identify six significant pillars for developing the framework as: Social, Management, Economic, Legal, Technology, and Sustainability (SMELTS) and explain how these factors can make the smart city project successful [45]. Dameri et al. examine several smart cities and observe that no consolidated standards or best practices for smart city governance are applied; instead, each city implements its own governance framework [46]. Yigitcanlar et al. discuss how drivers of smart cities—community, technology, and policy—construct the desired city outcomes: productivity, sustainability, accessibility, well-being, livability, and governance [47]. These studies show that even if the matter is one, the perspectives differ. It is clear that the authors encourage the development of a proper governance framework for IoT smart city applications, but their approaches are not the same. While some claim that there is no “one size fits all” solution, other groups believe that specific standards or pillars should be

attained. And within the latter, each group argues in favor of different pillars for smart cities. This shows the need for a unified framework model that is detailed enough to target aspects of the smart city system, but broad enough to be applied in most cities.

B. Self-driving Vehicles

Self-driving cars are a famous example of IoT applications. To be able to drive on its own, a self-driving car requires thorough analysis and processing of data that involves information about the road, traffic, and navigation. Kroger believes that transparency and communication should be central to the process of introducing self-driving cars [48]. Enlarged stakeholder involvement is advisable in a stepwise governance approach. Also, the public needs to be sensitized and better notified about expected rises and falls of advancements of selfdriving cars and mitigation methods should be verified. Bundin et al. discuss the need for consistent technical supervision of the operation of self-driving cars, that is the hardware and software aspects, besides the ethical issues of responsibility and autonomy of AI-powered drivers [49]. Also, they highlight the need to evolve the current road infrastructure to adapt to self-driving transport services. Cohen et al. see the role of government in the initiative of self-driving vehicles as risk mitigation, infrastructure adaptation, and preservation of public trust [50]. In their study, they notice how fast issues and concerns related to this technical advancement would come up. They conclude that approaches to governing innovative projects seem to face troubles quickly. Hanna et al. believe that self-driving vehicles are a sort of disruptive technology that will shift the federal legislative, regulation, development, and research [51]. Educating the public through well-informed and data-driven policies is needed to ensure privacy and industry-safe practices. These studies signify the disparity in perspectives regarding self-driving vehicles. However, a common topic among the studies is the need for adaptation to the IoT application while preserving public trust. This shows the importance of

ensuring proper and clear communication upon implementing IoT services, which will be discussed more thoroughly throughout this chapter.

C. Smart Farming

Rapid developments in IoT propelled the phenomenon of smart farming. Using Big Data, smart farming establishes management tasks based not only on location but also on data that is enhanced by context awareness and real-time events. Wolfert et al. expect Big Data to shift power and roles and heavily impact not only farming but the entire supply chain through smart sensors and IoT decision-making devices [52]. Thus, they believe that adequate smart farming governance, security, and business models are key matters to be addressed in future research. Boursianis et al. analyze the Unmanned Aerial Vehicle UAV technology and conclude that it, alongside IoT, is a major technology capable of transforming traditional farming practices into a new form of intelligence in precision agriculture [53]. Debauche et al. study how the absence of interoperability between smart farming platforms hinders the possibilities of advanced applications and innovative operations [54]. They recommend addressing security matters from the IoT design level. Just like in the previous IoT applications, researchers in the domain of smart farming tend to focus on general challenges and broad recommendations while advocating for a narrowed-down approach. This chapter aims to fulfill this goal and present solutions that can be tailored to different IoT systems.

3.3) Governance Frameworks for AI and IoT

Several governance frameworks and guiding standards currently exist to manage the applications and services of AI and IoT. While some governance strategies target the design level only, others are applied in the implementation stages. Table I summarizes governance guidelines across different phases and systems.

A. Governance Integrated into the Design Phase

Singapore's Infocom Media Development Authority IMDA IoT Cybersecurity Guide presents principles and respective recommendations depending on the design phase. It provides recommendations for IoT developers, solution architects, and system integrators who want to design, develop and deploy secure IoT products and systems. The recommendations cover four fundamental IoT security design principles: security, defense, accountability, and resiliency [55]. They emphasize the importance of employing strong cryptography, conducting threat modelling, enforcing proper access control, and more.

B. Governance Integrated into the Implementation Phase

We study the National Institute of Standards and Technology (NIST) guidelines [56]. These standards target the IoT system starting from evaluating the device identification and configuration to the level of data protection and cybersecurity state awareness. They also stress the importance of the regular need for authorized software and firmware updates and logical access to interfaces. Moreover, the Cloud Security Alliance (CSA) IoT Control Framework follows an extensive approach to manage the security challenges within IoT applications [57]. It covers baselevel security controls needed to diminish many of the risks resulting from the multiple networking technologies, cloud services, and connected devices incorporated in IoT systems. It encompasses,

in detail, several aspects of security such as: risk mitigation, technical policies, privacy, governance, secure connections and development, vulnerability, liability, misuse, physical security, secure vendor remote access, and end-of-life of devices. This allows the framework to help users manage highly sensitive systems supporting critical services, identify the proper security controls, and assign them to the appropriate IoT components [57].

Also, the Personal Data Protection Commission's (PDPC) Model AI Governance Framework provides private sector organizations with guidelines on how to deploy AI in their systems and address ethical challenges [40]. It aims to promote public understanding and trust in technologies by providing a clear explanation of how AI systems work, building adequate data accountability practices, and offering transparent communication. The guidelines are divided into principles that require AI systems to be human-centric and decisions to be explainable, transparent, and fair and practices that regulate the internal measures in the organization, the level of human involvement, operation management, and communication.

In addition, the Guide to Job Redesign in the Age of AI helps organizations and employees deal with the job-related challenges of AI; meaning it helps them understand how existing job roles can be redesigned to utilize the potential of AI, so that the work productivity is increased [58]. It provides a practical approach to help companies and organizations prepare for the digital future while managing AI's impact on employees. Its focus is human-centric, as it focuses on employees and re-skilling employees and redesigning jobs to match the market.

Furthermore, ISAGO - Implementation and Self-Assessment Guide for Organizations is a companion guide to the PDPC model framework mentioned earlier [59]. It helps organizations evaluate the compliance of their AI governance practices with the Model Framework. Organizations are encouraged to document and record the development of their governance

process as a matter of good practice, and for their experience to be used as case studies that would benefit other organizations and businesses. Organizations should not attempt to implement all the considerations in this Guide because not may be applicable in their context. Moreover, some considerations can only be applied in specific situations and cases.

C. Governance integrated in both Design and Implementation Phases

ISACA, formerly known as the Information Systems Audit and Control Association, is a professional and international association that focuses on risks, challenges, and management strategies of IT applications [60]. For IoT, it aims to evaluate security and control built in at design level, test controls and watch out for vulnerabilities, and educate and train individuals involved to ensure that building security alongside functionality. This is necessary for IoT security, in which experienced IT assurance and security personnel who understand cyber matters can evaluate IoT risks and benefits.

D. Explainable AI: SHAP and LIME for AI Governance

Explainable AI, often referred to as XAI, serves the purpose of clarifying an AI model, its anticipated effects, and the possible biases it might carry. Its role encompasses defining the model's precision, fairness, transparency, and the results it yields in AI-driven decision-making scenarios. The integration of XAI becomes paramount for organizations seeking to instill trust and assurance when deploying AI models for practical use. Additionally, AI explainability plays a pivotal role in assisting an organization in embracing a responsible stance toward AI development [61].

Phase	Guideline Summary
Design	IMDA IoT Cybersecurity Guide focuses on security, defense, accountability, and resiliency during the design phase.
Implementation	• NIST guidelines emphasize device identification, data protection, and the need for authorized updates. • CSA: IoT Control Framework addresses multiple aspects of IoT application security. • PDPC's Model AI Governance Framework offers guidelines for ethical AI deployment. • A Guide to Job Redesign focuses on human-centric AI's impact on job roles. • ISAGO aids organizations in AI governance practices.
Design & Implementation	ISACA focuses on risks and strategies of IT applications. It emphasizes building security alongside functionality for IoT.
XAI	Explainable AI (XAI) aims to clarify AI model behaviors. SHAP and LIME are post hoc analysis techniques that enhance model interpretability. They can be critical in studies, like smart city crime prediction, to ensure transparency, trust, and bias detection.

Table 1 Governance Frameworks across different phases and systems

Two of the XAI techniques that we'll introduce in this chapter are SHAP and LIME, which both use Python libraries. Within the Shapley Additive Explanation (SHAP) framework, influenced by a concept originating from game theory, the dispersion of predictions is allocated across the existing covariates. Consequently, it becomes possible to evaluate the impact of each explanatory variable on individual predictions independently of the underlying model. As for Local Interpretable Model-Agnostic Explanations (LIME), it is a technique for explaining machine learning models, regardless of their complexity, by creating a simplified, interpretable model specifically tailored to explain individual predictions. This approach is applied after the original model has made its predictions (as a 'post-hoc' step) and is designed to enhance our understanding of each prediction [62].

SHAP and LIME assume a critical role in post hoc analysis, enhancing the interpretability and explainability of models. These techniques come into play once the model has been crafted, trained, and implemented, offering valuable insights into the model's behavior and prediction rationale. Although they are not involved in the initial design and training stages, their importance

lies in facilitating a deeper comprehension of the model's efficacy and reliability, thereby aiding in its refinement.

To further elaborate on the contributions of SHAP and LIME to AI governance and policing, we will introduce how they can play a role using one of our previous smart city studies. In this study [63], we collect sensory traffic data, use machine learning tools to detect patterns within these data and perform predictions, and consequently develop traffic crime indicators that help with proactively enhancing road safety measures and overall protection standards. If we were to incorporate SHAP and LIME into our study, SHAP would reveal how the used models have arrived at their decisions by highlighting the importance of each different feature detected and LIME would provide insights into why a specific crime prediction was made depending on a particular case or location. Smart crime prediction is a sensitive research area, so this detailed analysis approach is critical for transparency and trust, bias detection and mitigation, and enabling effective communication with stakeholders. A thorough analysis of LIME and SHAP will be done in the next chapter.

3.4) Blended AI and IoT Governance Framework: A Discussion

It is clear that no governance approach is identical to the other. However, there are common aspects of interest between all. Johan Smith et al. find that there is not one proposed security framework that tackles all layers of IoT applications and development [64]. They also stress on the importance of addressing security from the IoT design phase, as it allows developers to attain a clearer understanding of the IoT system. From what has been mentioned, it is evident that governance methods are data-specific, the design phase in IoT development is crucial to monitor

the entire IoT application, data security is a must, and so is ensuring the security of the other social/digital applications that would assist in IoT security.

Moreover, research in IoT domain faces a lot of constraints, IoT governance board is needed to ensure compliance, and the ethical and social aspects play a major role in implementing successful IoT governance. If we study the matter deeper, we realize that the same comments can be applied to AI systems. This shows that it is possible to use AI governance frameworks and alter them to match IoT specifications. Researchers, technicians, analysts, and specialists can review previous AI guidelines and regulations and choose the most successful and relevant ones to implement in IoT systems.

As for XAI approaches, SHAP, LIME, and other tools focus on the constant challenge in AI and ML-enabled IoT studies, which is ensuring transparency, fairness, and clarity throughout the data collection and analysis processes. For individuals involved in decision-making and other relevant parties, these techniques provide a transparent insight into the reasoning behind AI-powered decisions, empowering them to make more knowledgeable decisions. Moreover, they aid in the detection of possible biases or model inaccuracies, ultimately resulting in outcomes that are both fairer and more precise [65]. However, as AI continues to advance and become more complex, understanding the algorithm to determine how and why the outcome was reached is becoming more challenging. Many explainable AI models involve simplifying the primary model, and this leads to a loss of predictive performance and consequently less accuracy. Moreover, existing explainability methods may not cover all the phases of the decision-making process, which can restrain the benefit of the explanation, especially when dealing with more complex models and more complex smart domains. Also, further advancement of the AI models will lead to the advancement of the explainability model, which will be a challenge for human analysts.

3.5) Pillars of an Ideal Blended AI-IoT Governance Framework:

In light of these considerations, it is evident there are several concepts that need to be included to achieve proper AI control, regulations, and decision-making. We can conclude that the ideal governance framework for AI-powered IoT systems should tackle (not limited to) 5 main notions:

• Data design and security: This regulates the creation, collection, storage, and management of the data across the different layers of the IoT reference architecture for it to meet the organization and customers' needs. It also makes sure that the system aligns with the communication standards and protocols of other digital systems at the organization.

• Operation and management: This dimension controls the integrated workflow within the digital system and with other IT systems. This is achieved through real-time insights and data analytics; the frequency and specificity of the IoT data indicate a high level of correlation control points, time-related policies to compensate for data loss, and data-trust checkpoints for data sources.

• Governance structure and communication: This governs the implementation process of AI/IoT and ensures proper communication with all parties involved in the process (Employees, Stakeholders, Customers). This is achieved by making sure that there is a motive to set up a governing committee and specify measures across departments. Hence, the organization board would support the AI governance structure, the assigned responsibilities of the personnel would be clear and well-defined, and a communication and explanation strategy (with backup plans) would be set up with a clear audience, context, and purpose. Also, the stakeholders would be informed of the impacts of AI, possible risks, and whether they are reversible or not. Moreover, the option to opt out would be applicable and a platform for feedback and queries would be managed by the appropriate personnel.

• Compliance with ethical and legal standards: As mentioned in the previous chapter, this ensures that the decision to use AI and IoT should comply with social, ethical, and legal considerations. This is done by making sure that expected benefits outweigh the costs, no human rights were ignored/overlooked/threatened at the time of design and usage of data, individual involvement was enough and appropriate, and there is room to motion to politically and officially alter traditional laws that might challenge the innovative aspect of IoT and AI.

• Transparency and interpretability: It is vital for an organization to have a full understanding of the AI decision-making processes with model monitoring and accountability and not to blindly trust them. Bias, often based on gender, race, or location, has been a long-standing risk in training AI systems. Moreover, the performance of AI models can degrade or drift because production data is different from training data. This makes it critical for a business to continuously manage and monitor models to promote AI explainability while measuring the business influence of using such algorithms. Explainable AI also helps promoting transparency, model auditability, end-user trust, and productive use of AI, and ensures the stakeholders understand the models' decision-making process. It also mitigates compliance, security, legal, and reputational risks of AI systems [65].

Our proposal revolves around a human-centered approach. Main standards include human involvement in decision-making, clarity of the AI deployment and understanding of the system procedures, accuracy and reliability of data sets, communication between staff and stakeholders, compliance of outcomes with the initially intended objectives, and ability of risk and bias management. If these standards were to be taken into consideration by tech workers and developers as a generic policy-making framework, then general and common IoT/AI risks would diminish. However, industry professionals should be aware that each environment has its own rules and plans and thus, they should carry their own actions accordingly. Therefore, depending only on one

conventional context will leave exceptional and system-specific risks unhandled. For that, our suggested policy should be taken as a checkpoint plan, where each company or organization should further elaborate on and subdivide each standard according to its specific requirements and qualities.

3.6) Summary

IoT applications have the ability to affect communities and change the lives of citizens. The impact, however, depends on how each IoT system is managed. This chapter reflects on previous research findings and model frameworks regarding the governance of IoT and AI and highlights the major factors that one should focus on from the design level to the implementation level of IoT and AI systems. We conclude that governance should be specific to each region, and a "one-size fits all" vision will not work. Moreover, the inclusion of all sectors: governments, the private sector, and academia, and AI tool, is a must to ensure a transparent and fair innovation. Transparency should not only include a clear explanation of the projects/plans, but also the expected rise and fall of the relevant IoT application. Also, while developing a framework, attention should be given to risk management when it comes to both aspects of assessment and mitigation. The integration of AI and IoT technologies bring huge benefits to the adopting systems. However, it remains a transformative approach with irreversible features. This stresses the importance of harnessing their potential and safeguarding against potential pitfalls and setting adequate management and supervision to create a transparent, safe, and trusted environment for all individuals directly and indirectly involved; employees, customers, and stake/share-holders [66].

Once transparency and explainability have been clarified, attention in the next chapter turns towards fostering fairness and equity within AI systems. This entails placing a significant emphasis

on identifying and addressing biases through detection and mitigation strategies. Furthermore, evaluating the effectiveness of these measures falls under the purview of XAI, ensuring that the decisions made by AI systems are not only transparent but also equitable and free from discriminatory biases.

4. Unveiling Fairness: XAI's Role in Bias Mitigation in Machine Learning Applications

4.1) Introduction

Many businesses now leverage AI, specifically machine learning ML to analyze vast amounts of data, covering tasks ranging from assessing credit for loan applications to examining legal contracts for mistakes and supervising employee interactions with customers to identify inappropriate behavior [67]. AI and ML algorithms learn from data to make predictions. However, research is uncovering instances where algorithmic decision-making falls short of expectations and unintentionally mirrors and intensifies existing biases, especially against vulnerable populations.

The preceding chapters discuss the risk of bias often and highlight the significance of governing strategies across various stages of AI-enabled IoT applications, particularly emphasizing the design and implementation phases, and suggest leveraging XAI to support the supervision. This chapter narrows its focus specifically on ensuring transparency and fairness within these phases. It ventures into practical applications, spotlighting key features and illuminating potential biases that might surface. Underscoring our dedication to transparency and the reduction of AI bias, we introduce "PrAIse," a pioneering tool designed to automate the explainability and bias detection process.

4.2) XAI, Fairness, and Bias: Background

The creation and use of machine-learning algorithms have grown easier for developers, making vast datasets readily available for computers to extract new insights. This has advanced algorithms into complex, widely-used tools for automating decisions.

4.2.1) The importance of and need for Explainable AI (XAI)

Explainable AI, commonly known as XAI, is generally a multidisciplinary area of research that is concerned with developing methods to make AI comprehensible to humans [68]. Its role encompasses defining the model's precision, transparency, fairness, and the outcomes it produces in AI-driven decision-making scenarios. The integration of XAI becomes crucial for organizations aiming to foster trust and confidence when deploying AI models for practical use [69].

Beyond moral considerations, legal and regulatory obligations mandate fairness and transparency in AI decision-making. The General Data Protection Regulation (GDPR) of the European Union, for instance, emphasizes the right to explanation, ensuring individuals can comprehend the reasoning behind automated decisions affecting them [70]. Similar legal systems and regulatory bodies are increasingly stressing the importance of explainability, underscoring the need for XAI systems to provide clear justifications for the outcomes of AI and ML predictions.

Moreover, explainable AI can assist in identifying biases and discrimination within AI systems, allowing for the mitigation of unfair outcomes. Biases in training data or algorithmic processes can lead to discriminatory decisions, perpetuating societal inequalities. XAI techniques can

illuminate the factors contributing to these biases, enabling corrective actions to ensure fairness and equal treatment from the beginning [71].

This work focuses on two XAI techniques, namely Shapley Additive Explanations “SHAP” and Local Interpretable Model-agnostic Explanations “LIME”, introduced in the previous chapter. While the former enables the assessment of each explanatory variable's impact on individual predictions independently of the underlying model, the latter is a method for elucidating machine learning models, irrespective of their complexity, by constructing a simplified, interpretable model tailored to elucidate individual predictions. This strategy is implemented after the original model has made its predictions, aiming to enrich our comprehension of each prediction.

4.2.2) A comprehensive examination of Fairness and Bias

Fairness is a confusing concept [72]. Fairness encompasses multiple definitions. One source defines it as the absence of favoritism towards any side [73]. Another defines it as the quality of treating people equally or reasonably [74]. However, the concept of fairness varies depending on context and perspective. In law, it involves protecting individuals and groups from discrimination, while in social science, it considers fairness within the framework of social relationships and power dynamics [75]. Philosophically, fairness is often linked to moral rightness and notions of justice and equity. In the realm of AI and ML, fairness is a pervasive concept, with a fair model aiming to achieve "equalized odds," ensuring group membership has no direct effect on an individual's score [76]. In essence, bias and fairness in AI are interconnected, with fairness broadly defined as the absence of prejudice or preference based on

characteristics [77]. Just like a taxonomy for fairness definitions is established by different researchers across different domains, there exists multiple classification of bias.

In our study, we focus on Bacelar's bias level classification model, which consists of 4 types (figure 2), for the sole purpose of describing the existing different levels of bias [78]:

Bias in AI and machine learning can manifest in several ways:

- **Institutionalized Bias**: This type originates from the training data, reflecting historical and societal inequities or human decisions. For example, word embeddings from news articles might perpetuate gender stereotypes. Bias can also arise from data collection methods, such as crime records inflated by over-policing in certain areas, leading to a cycle of biased data.
- **Individual Bias**: Human prejudices infiltrate the data and the analytical process. The biases of scientists can influence the analysis, experiments, or implementation of algorithms, with overlooked parameters or other oversights potentially leading to skewed outcomes.
- **Sample Bias**: This occurs when the training data sample isn't representative of the whole population, often due to systemic collection errors. Algorithms trained on such biased samples can unfairly favor or disadvantage certain groups.
- **Algorithm Development Bias**: Bias can be introduced during the creation of algorithms through numerous decisions, such as dataset selection, feature encoding, and the choice of outcome variables. These decisions are based on implicit, often normative assumptions, hidden behind the algorithms' mathematical and procedural precision.

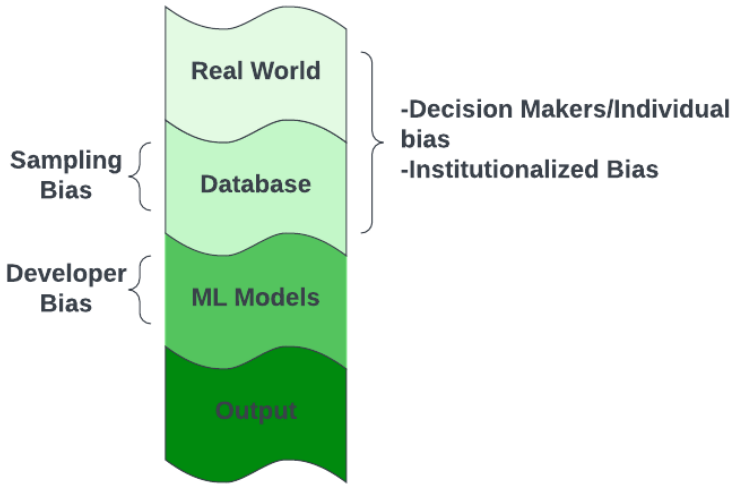

Figure 2 - Bias Types in Different Phases

In light of the significance of XAI in addressing bias and promoting fairness, it is crucial to examine the relevant research on incorporating XAI into bias mitigation strategies.

4.2.3) Relevant Work in the Area

Slack *et al* and Dieber *et al* contribute significantly by directing their focus towards the inherent limitations of LIME and SHAP [79] [80]. Their study extends beyond mere identification of limitations, delving into a demonstration of how specific biased classifiers adeptly manipulate these popular explanation techniques. Their findings show how biased classifiers strategically

deceive LIME and SHAP, effectively generating explanations that superficially appear inoffensive while concealing the latent biases within the models.

In the work of Arias-Duart *et al.,* they introduce an evaluation metric, "Focus," for assessing feature attribution methods in image recognition [81]. These methods aim to explain a model's behavior using visual cues. Additionally, the paper presents a novel methodology for incorporating noise from within the distribution, as opposed to using out-of-distribution noise as done in previous studies.

Bertrand *et al* propose a heuristic map linking human cognitive biases with explainability techniques in eXplainable AI (XAI) [82]. This map is based on a systematic review of 37 papers, selected from a corpus of 285, which highlight cognitive biases in XAI systems. The review reveals four primary interactions between cognitive biases and XAI: 1) influencing the design of XAI methods, 2) distorting evaluations of XAI techniques in user studies, 3) potential mitigation of some biases by XAI, and 4) potential exacerbation of other biases by XAI.

Sufian and Bogliolo propose a fairness-aware mechanism for user-friendly explanatory systems, integrating gradual learning to retrain the machine learning model based on counterfactual explanations during online interaction [83]. Their experiment examines bias and bias mitigation in interactive explanatory systems. Their contribution involves processing new data for evaluation, comparing it with old data for class imbalance and drift, and introducing an incremental learning-based ensemble of ML models to enhance performance for class-label prediction of new data points.

The literature highlights the promise of XAI methods, such as LIME and SHAP, in making AI models more transparent. Yet, it also points out their vulnerabilities, The findings advocate for

creating stronger, bias-conscious explanatory systems that clarify AI decisions and help lessen biases, stressing the need for automated explanation and bias identification to make AI more transparent and fairer. Building upon these insights, our study conducts an experimental evaluation to assess the effectiveness of XAI in conjunction with other bias-detecting features. Our work stands out by evaluating eight machine learning models across two diverse use cases, emphasizing both transparency and fairness. Unlike others focused on a singular aspect, we examine both in varied contexts. Additionally, we go beyond standard methodologies by showing how "PrAIse," our automated tool, can provide a quicker, more efficient means to achieve our goals, marking a significant advancement in ethical AI practices.

4.3) Methodology

In our methodology overview, we embark on a rigorous examination of machine learning models across two distinct use cases: credit defaulting and recruitment. For each, we meticulously train a suite of diverse ML models to generate predictions, subsequently scrutinizing these predictions for accuracy, precision, and overall performance. To unravel the intricate logic of our models' decisions, we employ SHAP and LIME—two cutting-edge interpretability techniques. SHAP grants us the power to evaluate the influence of each explanatory variable on individual predictions, independent of the model's complexity. Meanwhile, LIME offers us a window into the model's reasoning by creating a simplified, interpretable surrogate model that captures the essence of individual predictions.

Our commitment to fairness is unwavering, and we evaluate it using a set of critical metrics: accuracy, equal opportunity1 (true positive rate: TP / (TP + FN)), and equalized odds (false positive rate: FP / (FP + TN)).

It is crucial to recognize that the interpretation of bias metrics can vary significantly across different environments and scenarios and is often contingent on the specific definitions of fairness pertinent to each context. Accuracy, Equal Opportunity, and Equalized Odds, in this study, are interpreted based on their definitions and the specific fairness considerations they address.

To streamline and automate this analytical rigor, we introduce "PrAIse," an innovative tool engineered for explainability and bias detection. PrAIse ensures transparency and fairness are at the forefront, enabling users to apply the principles of ethical AI seamlessly within their models and datasets.

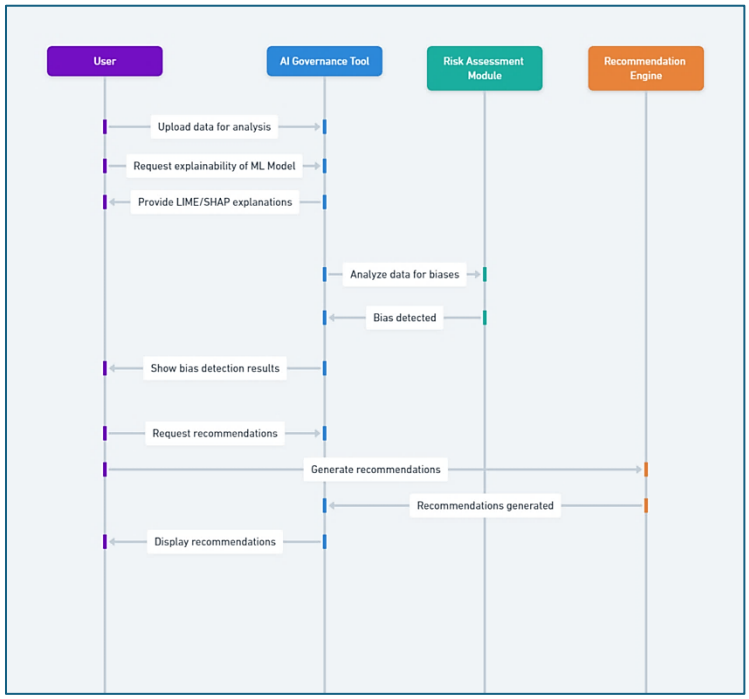

Figure 3 - PrAIse Process Diagram

Figure 3 shows how PrAIse tool integrates model explainability and bias detection into a cohesive workflow. Initially, users can upload their data for analysis.

The tool then employs model Explainability techniques to deconstruct the AI model's decision process. This step generates insights into which features are most influential in the model's predictions, providing transparency.

Following the explainability analysis, the tool proceeds to bias detection. It evaluates the model's decisions for systematic favoritism or prejudice against certain groups, based on the relevant features. Detected biases are then assessed for risk levels, which are classified and displayed to the user. Based on this, the tool suggests actionable recommendations to enhance model fairness and help users select mechanisms to enforce governance process within their use cases.

In the following sections, we will provide an overview of the outputs that would be presented by the tool for each use case and explain how these insights can be leveraged for practical benefit.

4.4) Evaluation

4.4.1) Recruitment Use Case

A thorough qualitative analysis has unveiled the presence of bias within recruitment and hiring processes, raising concerns about adherence to established ethical guidelines and nondiscrimination policies. The Universal Declaration of Human Rights (UDHR) Article 23 underscores every individual's right to work and to a free choice of employment without discrimination, guaranteeing equal pay for equal work [84]. Similarly, the International Labour Organisation (ILO) defines discrimination in employment as any distinction, exclusion, or preference based on race, color, sex, religion, or other factors that hamper equality of opportunity or treatment [85].

Despite these clear directives, empirical evidence suggests these principles are not universally upheld. A harmonized comparative field experiment spanning six countries revealed systemic

bias against male applicants in Germany, the Netherlands, Spain, and the UK, and consistently found female applicants favored for roles in female-dominated sectors, with no parallel preference for male applicants [86]. Nevertheless, three female employees from Amazon's corporate research and strategy division filed a lawsuit against the company, alleging gender discrimination and retaliation [87]. They accuse Amazon of consistently giving female staff lower job titles compared to men who are in the same roles but with higher titles and salaries. Moreover, Furthermore, Quillian et al. concluded that discrimination in hiring is less prevalent for jobs that require a college degree compared to those requiring only a high school diploma or equivalent [88]. These results show a clear gap between what international rules recommend and what happens in job hiring, emphasizing the widespread difficulty in reaching real equality at work.

Human capital studies show that over 50% of employers are expected to be using AI for recruitment [89]. Therefore, we will assess and evaluate ML-aided recruitment approaches, using a publicly available dataset from Kaggle, specifically tailored for HR analytics and understanding the factors influencing data scientists' job changes and potential recruitment [90].

a- **Dataset Preprocessing**: Prior to modeling, we conducted thorough preprocessing of the dataset (Table 2) to ensure optimal quality of input data. This involved handling missing values, outlier, and null values within several features including “Gender”, “Education Level”, “Experience”, and “Company Type”. The approach for handling null values was adopted to create meaningful categories and reduce feature complexity for the models.

Decisions to drop, bin, or impute values were made based on the nature of the data and the impact on model predictions.

Feature Name	Description
enrollee_id	Unique ID for candidate
city	City code
city_development_index	Development index of the city (scaled)
gender	Gender of candidate
relevant_experience	Relevant experience of candidate
enrolled_university	Type of University course enrolled if any
education_level	Education level of candidate
major_discipline	Education major discipline of candidate
experience	Candidate total experience in years
company_size	No of employees in current employer's company
company_type	Type of current employer
last_new_job	Difference in years between previous job and current job
training_hours	Training hours completed
target	0 – Not looking for job change, 1 – Looking for a job change

Table 2 - Feature Description of Recruitment Dataset

b- **Model Training**: Our focus is on comprehending both existing and new models, and evaluating them for biases, which is why we have adopted the steps in RecruitmentTargetML to train different machine learning models to predict the likelihood of candidates seeking a

job change, thus enabling the company to proactively target potential recruits [91]. For this reason, we train the model using Decision Tree Classifier, Random Forest Classifier, Logistic Regression CV, K-Neighbors Classifier (with n_neighbors = 7 and 101), SVC (Support Vector Classifier), Bagged Model (Decision Tree + KNN using bootstrap aggregating).

c- **Model Evaluation and Performance**: The accuracy and performance of each model were tested using standard metrics, including accuracy, precision, and recall (Table 1). The goal was to identify models that not only accurately predict job-seeking behavior but also maintain a high level of precision and recall to minimize false positives and negatives.

Model	Accuracy and Performance Measures		
	Accuracy	Precision	Recall
Decision Tree Classifier	0.773	0.548	0.403
Random Forest Classifier	0.783	0.574	0.430
Logistic Regression CV	0.776	0.591	0.268
K-Neighbors Classifier (n_neighbors = 7)	0.761	0.515	0.378
K-Neighbors Classifier (n_neighbors = 101)	0.780	0.576	0.386
SVC	0.776	0.557	0.418
Bagged Model	0.778	0.562	0.425

Table 3 - Accuracy and Performance Measures

- Looking at Table 3, our key findings are summarized as follows:

Random Forest Classifier shows the highest accuracy (0.783) among the models, with relatively balanced precision (0.575) and recall (0.431), indicating it's effective in predicting recruitment outcomes accurately while maintaining a good balance between identifying positive cases and avoiding false positives.

- Logistic Regression CV has a notable precision (0.592) but the lowest recall (0.269), suggesting it's cautious in predicting positive cases (e.g., a candidate being a suitable hire) but misses a significant number of actual positive cases.
- K-Neighbors Classifier with n_neighbors = 7 and n_neighbors = 101 both show a compromise between accuracy, precision, and recall, with slightly better performance in precision for n_neighbors = 101 (0.577). The variation illustrates the sensitivity of model outcomes to hyperparameter choices, emphasizing the importance of tuning in achieving optimal results.
- SVC and Bagged Model (Decision Tree + KNN) present similar accuracy levels (around 0.777 for both) with moderate precision and recall, indicating they are competent but not outstanding in any particular metric.
- Precision is generally higher than recall across models, suggesting a tendency towards conservatism in predicting positive outcomes. This might reflect a preference for minimizing false positives (e.g., wrongly predicting a candidate as a suitable hire), possibly at the cost of missing some true positives.

d) **Model Explainability**: To understand the reasoning behind the models' predictions and ensure transparency, we apply LIME and SHAP.

i) LIME Results:

LIME plots (Fig. 4 – Fig. 10) display the impact of each feature on an individual prediction made by a machine learning model. Positive and negative influences are shown with bars pointing to the right or left, respectively, indicating whether a feature makes the predicted outcome more or less likely. These plots provide an interpretable snapshot of how the model arrives at its decisions for a specific instance.

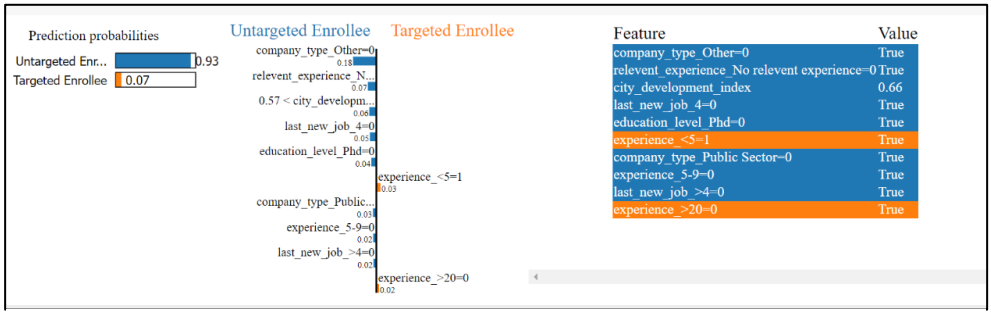

Figure 4 – Decision Tree Classifier Lime Results

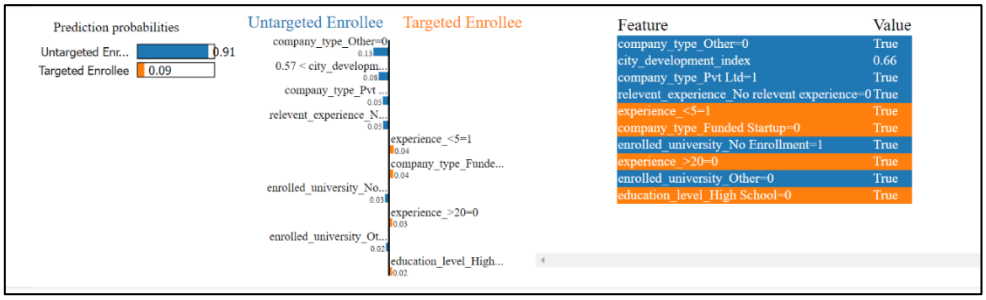

Figure 5 – Random Forest Classifier Lime Results

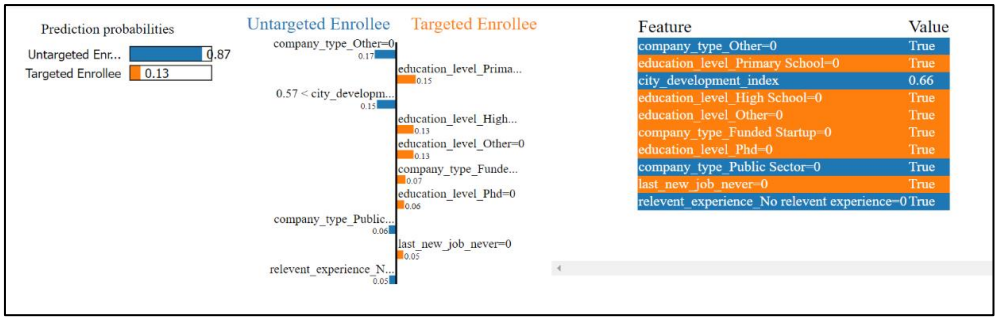

Figure 6 – Logistic Regression CV Lime Results

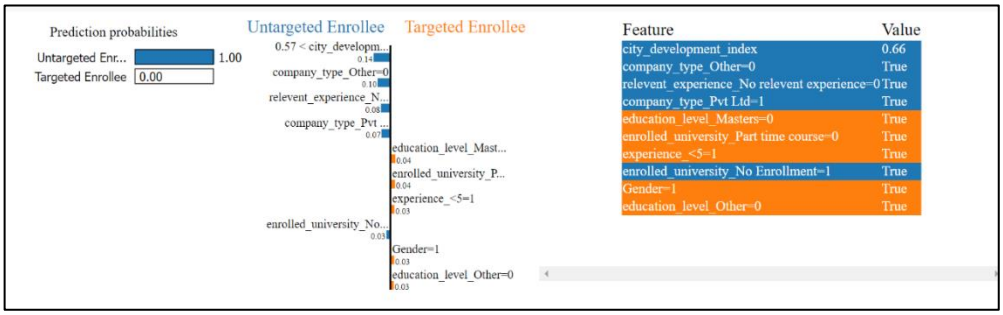

Figure 7 – K-Neighbors Classifier (n_neighbors = 7) Lime Results

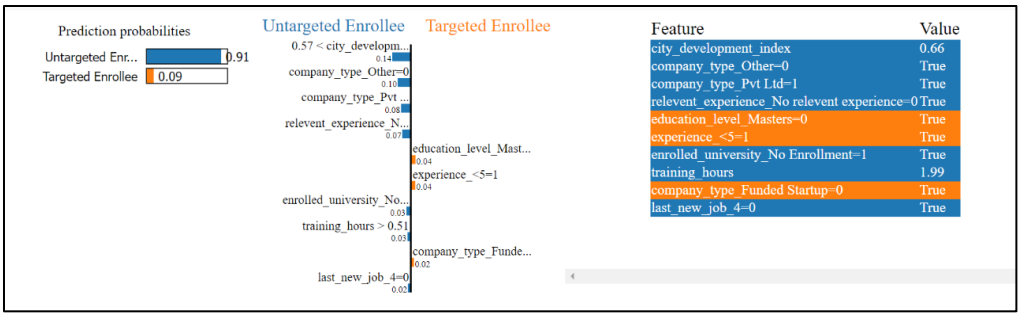

Figure 8 – K-Neighbors Classifier (n_neighbors = 101) Lime Results

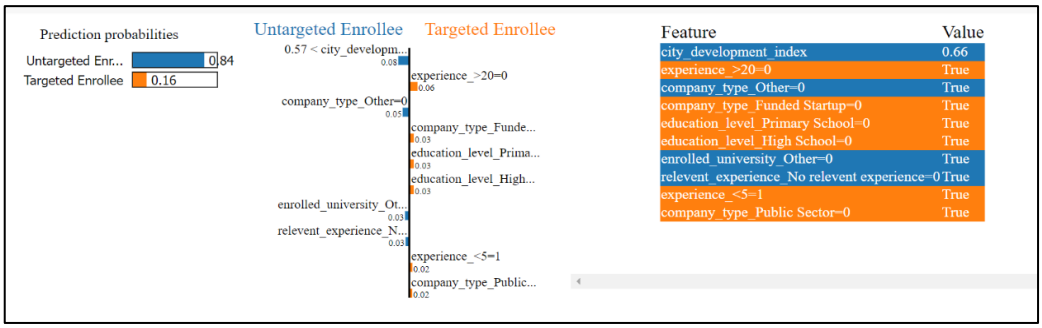

Figure 9 - SVC Lime Results

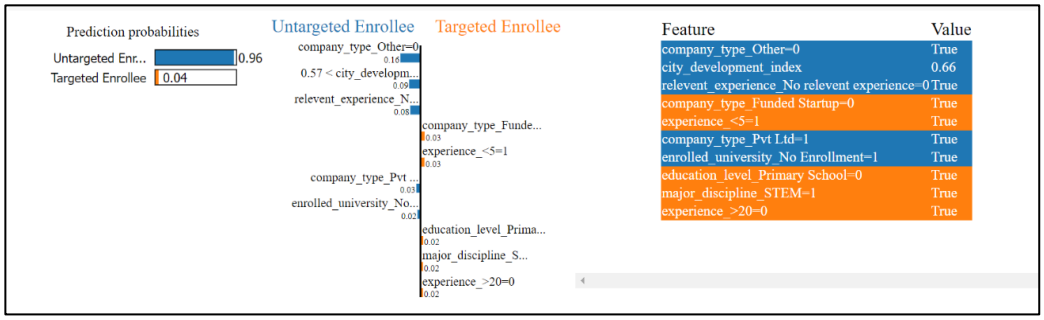

Figure 10 - Bagged Model (Decision Tree + KNN)

The LIME analysis plots reveal consistent themes in predicting 'Untargeted Enrollee' status across several models. Key findings include:

- 'City_development_index' and 'company_type' frequently emerge as pivotal in the decision-making process of being targeted for enrollment, likely reflecting the model's reliance on geographic or economic factors.
- Conversely, 'education_level' and 'relevant_experience' contribute differently in various instances; 'education_level_PhD=0', 'company_type_Public Sector=0', and specific 'education levels' often predict against being a 'Targeted Enrollee'.
- Despite varying feature influences, a strong tendency exists across models to predict individuals as 'Untargeted Enrollees', with probabilities often exceeding 0.90.
- This recurring prediction pattern suggests possible dataset class imbalance. Moreover, conducting such an analysis plays a crucial role in facilitating the identification and understanding of key bias metrics.

ii) SHAP Results:

In these plots (Fig. 11 – Fig. 17), each point represents a feature's SHAP value for an instance. Features pushing the prediction higher are shown in pink, and those pushing the prediction lower are in blue. The position on the x-axis indicates the impact magnitude, and the color represents the feature value (pink for high, blue for low).

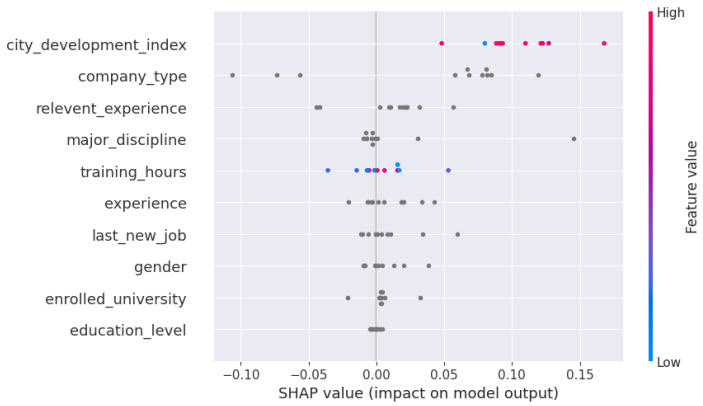

Figure 11 – Decision Tree Classifier SHAP Plot

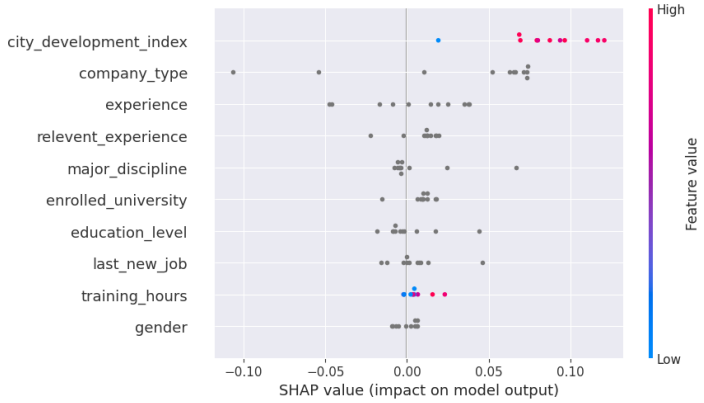

Figure 12 – Random Forest Classifier SHAP Plot

Figure 13 – Logistic Regression CV SHAP Plot

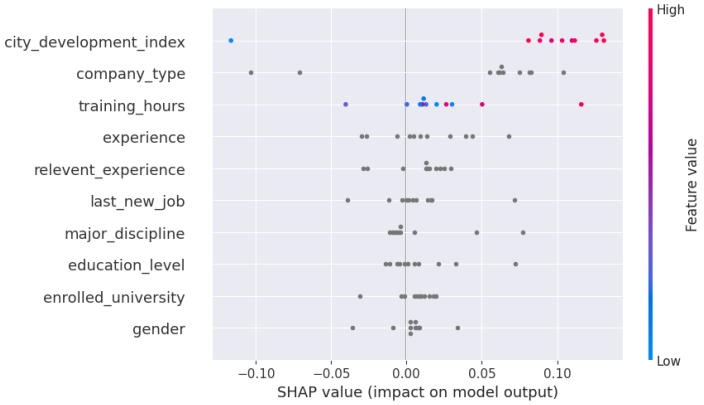

Figure 14 – K-Neighbors Classifier (n_neighbors = 7) SHAP Plot

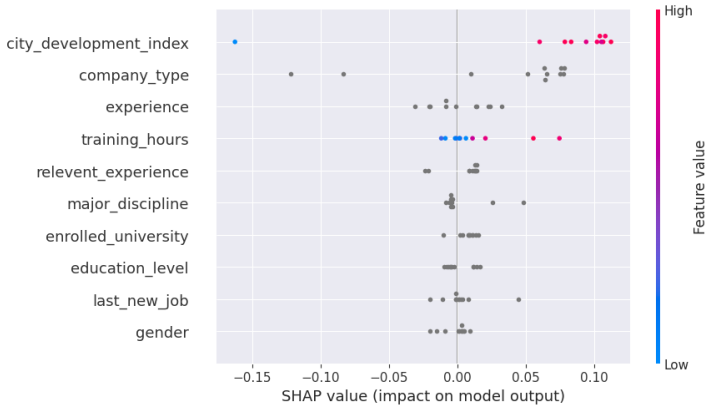

Figure 15 – K-Neighbors Classifier (n_neighbors = 101) SHAP Plot

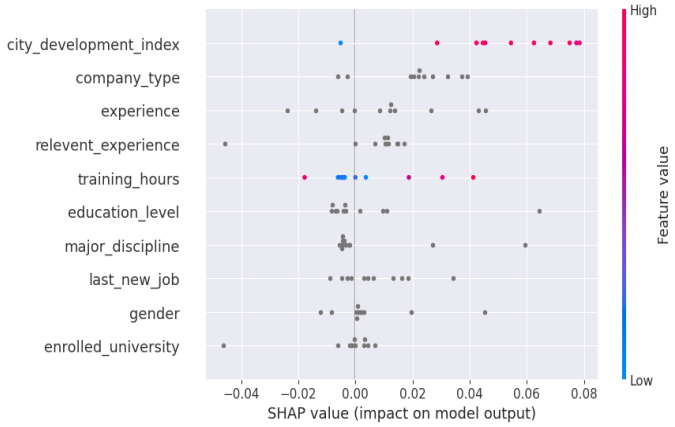

Figure 16 -SVC SHAP Plot

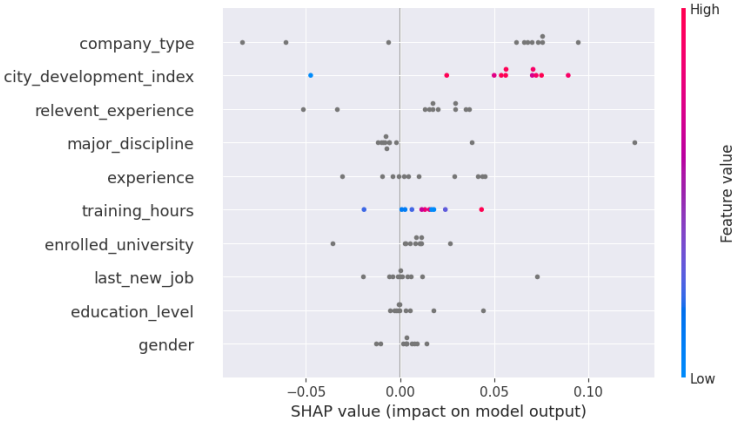

Figure 17 - Bagged Model (Decision Tree + KNN) SHAP Plot

Our key findings from the plots can be summarized as follows:

- ‘City_Development_Index’ is often a strong positive contributor to the model’s output, suggesting that candidates from more developed cities are more likely to be predicted as looking for a job change.
- ‘Company_type’ and ‘relevant_experience’ are also influential, but their impact varies across different instances.

- ‘Gender’ appears to have a less consistent impact across different predictions, which may indicate that it's not a primary factor in the model's decision-making process.
- The plots show a mix of positive and negative contributions from features like ‘education_level’, ‘major _discipline’, and ‘training_hours’, which suggests that their influence on the prediction changes depending on the specific candidate's profile.
- For some candidates, features like ‘last_new_job’ and ‘enrolled _university’ are negatively impacting the prediction, potentially implying these candidates are less likely to be looking for a job change.

Overall, these SHAP plots suggest that the model is most sensitive to geographical development and candidates' current roles. However, the influence of personal attributes like education and experience varies, indicating a more complex interaction between these features and job change likelihood.

e) **Bias Assessment:** To evaluate fairness in our predictive modeling, we scrutinize the bias metrics and examine fairness indicators, such as accuracy, equal opportunity, and equalized odds across the used machine learning models, to uncover any potential biases in predictions related to gender and educational background. For this task we use the following binary legends

   - *Gender =1 if Male*
   - *Gender = 0 if Non-Male*
   - *Edu level =1 if Graduate (Masters and PhD)*
   - *Edu Level= 0 if Other*

- *Positive Outcome: Individual Targeted for Employment*

Gender Key Findings (Table 4):

- The accuracy for Gender=1 (Male) is slightly higher than for Gender=0 (Non-Male) across all models, but the difference is minimal, suggesting similar predictive accuracy for both genders.
- LogisticRegressionCV shows a ratio of 0.187 for Gender=1 and 0.348 for Gender=0 under the Equal Opportunity1 metric. The ratio for Gender=1 is less than 1, indicating that males (Gender=1) are less likely than non-males (Gender=0) to receive a positive prediction when the true outcome is positive.

Bias Metric / ML Model	Accuracy			Equal Opportunity1			Equalized Odds		
	Gender =1	Gender =0	Ratio	Gender =1	Gender =0	Ratio	Gender =1	Gender =0	Ratio
Decision Tree Classifier	0.784	0.751	1.043	0.408	0.496	0.824	0.105	0.149	0.702
Random Forest Classifier	0.795	0.757	1.05	0.444	0.516	0.861	0.101	0.149	0.676
Logistic Regression CV	0.786	0.753	1.044	0.187	0.348	0.539	0.036	0.09	0.401
K-Neighbors Classifier	0.773	0.736	1.05	0.358	0.471	0.761	0.103	0.161	0.645
K-Neighbors Classifier (n_neighbors = 101)	0.789	0.758	1.041	0.334	0.457	0.732	0.076	0.125	0.606
SVC	0.793	0.757	1.047	0.382	0.498	0.768	0.086	0.142	0.603
Bagged Model	0.789	0.758	1.042	0.411	0.508	0.808	0.098	0.145	0.676

Table 4 - Bias Results for Gender

- Equalized Odds measure both true positive and false positive rates. Here, ratios are mostly below 1 for Gender=1 across models, suggesting a potential bias where males are less likely to be correctly predicted or more likely to be falsely predicted as the positive class compared to non-males.

Education Level Key Findings (Table 5):

- Accuracy is generally lower for individuals with graduate-level education (Masters or PhD) compared to others, indicating better model performance for Edu=0.
- Equal Opportunity1 ratios exceed 1 for Edu=1 across all models, suggesting that graduates are more likely to receive positive outcomes when they are due.
- Equalized Odds ratios are significantly above 1 for Edu=1 in all models, indicating a higher likelihood for correct positive predictions and a lower likelihood for false positives for graduates compared to non-graduates.

Bias Metric / ML Model	Accuracy			Equal Opportunity1			Equalized Odds		
	Edu=1	Edu=0	Ratio	Edu=1	Edu=0	Ratio	Edu=1	Edu=0	Ratio
Decision Tree Classifier	0.757	0.8	0.946	0.488	0.333	1.463	0.141	0.084	1.677
Random Forest Classifier	0.765	0.813	0.941	0.533	0.331	1.611	0.148	0.068	2.174
Logistic Regression CV	0.754	0.81	0.931	0.284	0.153	1.857	0.069	0.027	2.527
K-Neighbors Classifier	0.74	0.798	0.927	0.453	0.277	1.636	0.153	0.074	2.073
K-Neighbors Classifier (n_neighbors = 101)	0.761	0.808	0.942	0.446	0.227	1.96	0.12	0.048	2.503
SVC	0.765	0.808	0.946	0.507	0.239	2.125	0.138	0.051	2.721
Bagged Model	0.762	0.808	0.942	0.499	0.327	1.528	0.14	0.072	1.927

Table 5 - Bias Results for Education Level

f) PrAIse Analysis for Recruitment Scenario:

For the recruitment use case, the tool would analyze hiring algorithms for biases against gender and education level. Multiple machine learning models would show a trend where candidates from specific educational institutions or of a particular gender are consistently favored or penalized. In that case, the tool's explainability analysis would highlight these patterns, attributing decision weights to these factors.

After detecting such biases, the tool would suggest several remediation strategies. In our case study, such suggestions would be:

1. Rebalancing the dataset to ensure gender and educational background are represented proportionally to avoid class imbalance.
2. Applying algorithmic techniques to reduce the weight given to potentially biased features.
3. Encouraging diversity in training data by including candidates from a wider array of educational backgrounds.
4. Implementing fairness constraints or regularization within the model training process to explicitly penalize and therefore reduce bias.

4.4.2) Credit Default Use Case

After performing a comprehensive analysis, it is evident that bias can be easily present within the domain of banking, crediting, and financing, indicating a departure from the equitable practices advocated by global financial standards. According to the World Bank and the Federal Reserve

(the FED), Consumer Service Providers (CSPs) are mandated to make credit available equally to all creditworthy customers, strictly prohibiting discrimination [92]. Furthermore, FED Regulation B enforces a ban against discrimination on any prohibited basis, such as race, national origin, age, or gender, in credit applications, including the eradication of prohibited biases in credit scoring.

Despite these guidelines, evidence suggests that some institutions have not adhered to these principles as required, leading to discriminatory practices and bias in their application. For instance, a complaint by the US Justice Department against Patriot Mortgage highlighted that from 2015 to at least 2020, the company deliberately refrained from offering mortgage services to predominantly Black and Hispanic neighborhoods in Memphis and discouraged individuals from these communities from seeking home loans [93]. Additionally, a prominent software developer in the US shared his wife's experience of being denied an Apple Card credit line increase, despite her superior credit score and favorable factors, a predicament echoed by multiple women [94]. Furthermore, data from the World Values Survey, encompassing 57 countries, indicated that 60% of respondents felt that older individuals do not receive the respect they deserve [95]. This survey found negative attitudes towards older people, illustrating that such attitudes can permeate various sectors, including the realms of credit defaulting and financial services. This evidence underscores the ongoing challenges in ensuring fairness and equality in financial decision-making processes.

Just like the previous use case, we examine the predictive power of multiple machine learning models, assess their performance measures, explore the reasoning behind the predictions, and inspect the presence of possible biases regarding age and sex. Addressing these aspects is pivotal

for financial institutions to maintain bias-free assessments of customer creditworthiness, set appropriate credit limits, and manage risk effectively. Utilizing a dataset from the UCI Machine Learning Repository, and following the steps of Credit Default Analysis use case, our analysis delves into various aspects of credit defaulting, including monthly bill statements, payment statuses, and demographic information [96] [97].

a) **Dataset Preprocessing**: Before straining the models, we performed data preprocessing to make sure the data (Table 6) was of the best quality. This step involved addressing missing values in several features, such as “Limit Balance”, “Sex”, “Education”, “Bill Amount”, “Age”.

Feature Name	Description
ID (Unnamed)	Unique ID for customer
LIMIT_BAL (X1)	Amount of given credit in NTD, including individual and family credit
SEX (X2)	Gender of the customer
EDUCATION (X3)	Level of education
MARRIAGE (X4)	Marital status of the customer
AGE (X5)	Age of the customer
PAY_# (X6-X11)	History of payment behavior over six months
BILL_AMT# (X12-X17)	Amount of bill statement over six months
PAY_AMT# (X18-X23)	Amount of previous payments over six months
Default payment next month	Whether the customer defaulted the next month

Table 6 - Feature Description of Credit Default Dataset

b) **Model Training:** As we are dedicated to understanding and assessing both established and new models for biases, we train the following models: Random Forest Classifier, Logistic Regression, K-Neighbors Classifier, Dummy Classifier (Baseline Classifier/ Null Model), and SVC.

c) **Model Evaluation and Performance:** We tested each model's accuracy and performance with standard measures like accuracy, precision, and recall (see Table 7). Our aim was to find models that predict Credit-Defaulting accurately while also being precise and minimizing errors.

Model	Accuracy and Performance Measures		
	Accuracy	Precision	Recall
Random Forest Classifier	0.812	0.642	0.358
Logistic Regression	0.809	0.725	0.235
K-Neighbors Classifier	0.792	0.56	0.320
Dummy Classifier	0.654	0.215	0.208
SVC	0.782	0.726	0.038

Table 7 - Accuracy and Precision Measures

Our key findings can be summarized as follows:

- Random Forest Classifier: With an accuracy of approximately 81.26%, it has a balanced precision and recall. This suggests that while the model is reasonably accurate overall,

there may be some trade-off between precision and recall, which is typical in classification tasks.

- Logistic Regression: It has a slightly lower accuracy of about 80.96% but a higher precision score, which indicates it's better at predicting the positive class (credit will be defaulted), but tends to miss some positive instances (lower recall).
- K-neighbors Classifier: It shows slightly lower accuracy and precision than the Random Forest Classifier. Its recall is comparable to that of Random Forest, indicating it has a similar rate of true positive predictions.
- Dummy Classifier: This model has significantly lower scores across all metrics, which is expected since it should represent a baseline that makes predictions without any intelligent decision-making process.
- SVC: It has a similar accuracy to Logistic Regression and a high precision, but the recall is very low, indicating that while it's quite confident about the positive predictions it makes, it misses many actual positive cases.

d) Model Explainability: To understand the reasoning behind the models' predictions and ensure transparency, we apply LIME and SHAP.

i) LIME Results

LIME plots (Fig. 18 – Fig. 22) provide explanations for individual predictions, highlighting which features were most influential for the model's decision and in what direction.

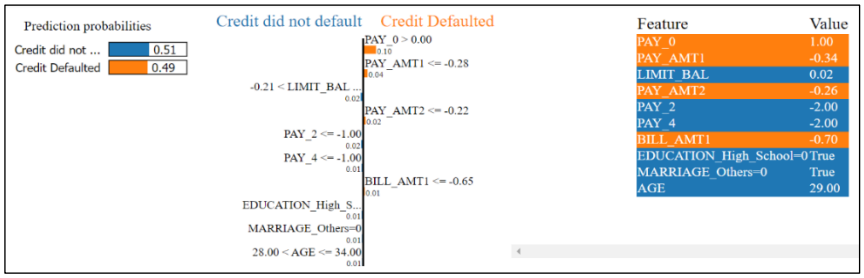

Figure 18 - Random Forest Classifier LIME Results

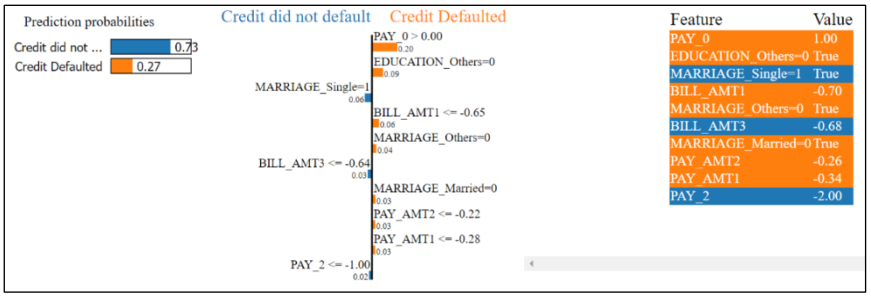

Figure 19 - Logistic Regression LIME Results

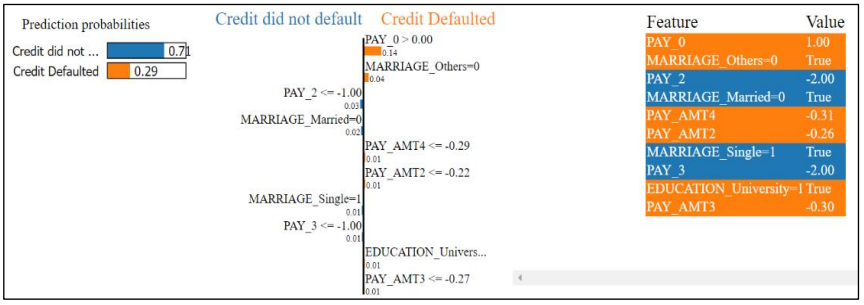

Figure 20 - KNeighbors Classifier LIME Results

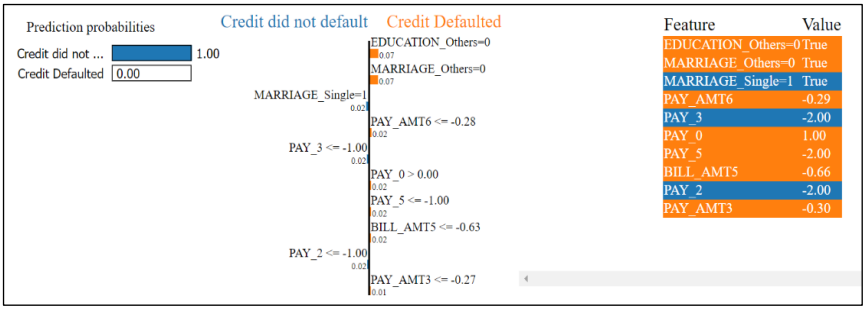

Figure 21 - Dummy Classifier LIME Results

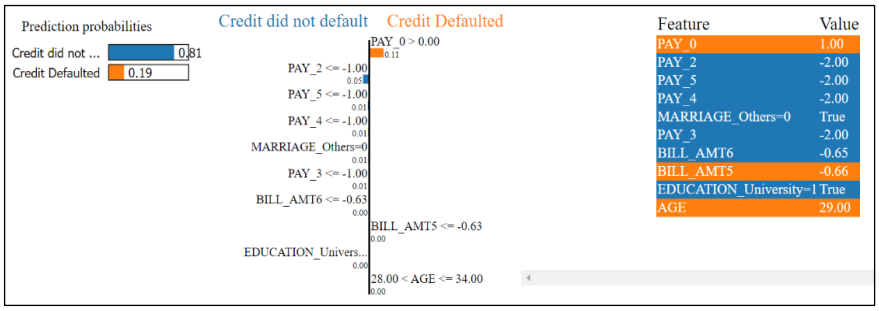

Figure 22- SVC LIME Results

The key findings of LIME Results are shown below:

- Recent Payment Status (PAY_0): This feature consistently has a strong positive impact on the probability of credit default across multiple instances. When PAY_0 is positive, it suggests that recent delinquency is a critical factor leading the model to predict a higher risk of default.

- Repayment Amounts (PAY_AMTX): The amounts paid in previous periods have varied influences on the prediction. In some instances, they contribute to a lower probability of default, indicating that higher repayments might be seen as a sign of better creditworthiness.

- Bill Amounts (BILL_AMTX): Similarly to repayment amounts, the bill amounts also show varied impacts. Lower bill amounts seem to contribute to a higher likelihood of credit default in some cases, which may indicate the model associates lower bills with financial distress or lower utilization of credit lines.

- Demographics (EDUCATION, MARRIAGE): Demographic features occasionally appear in the explanations with smaller contributions. For example, being single or having a certain education level may slightly impact the model's predictions, although these influences are less consistent and weaker than those of financial history.
- Credit Limit (LIMIT_BAL): In some explanations, the credit limit shows a slight negative impact on the probability of default, suggesting that higher credit limits might be associated with lower default risk.
- Consistency and Variability: While the influence of recent payment status is quite consistent, other features like repayment and bill amounts show more variability in their impact. Demographic features generally have smaller and less consistent effects.

The overall takeaway from the LIME explanations is that the model places considerable emphasis on recent payment behavior when assessing the risk of default. Financial behaviors, particularly recent ones, seem to be the primary drivers of the model's decisions, with demographic characteristics playing a secondary role. Each explanation provides insight into the model's behavior and identifies areas where it might be relying on unexpected or undesirable factors, which can allow decision makers not only to understand but also adjust models if needed.

ii) SHAP Results:

With SHAP, we gain a comprehensive understanding of feature importance and the magnitude of their effects on the predictive outcomes for credit default risk (Fig. 23 – Fig.27).

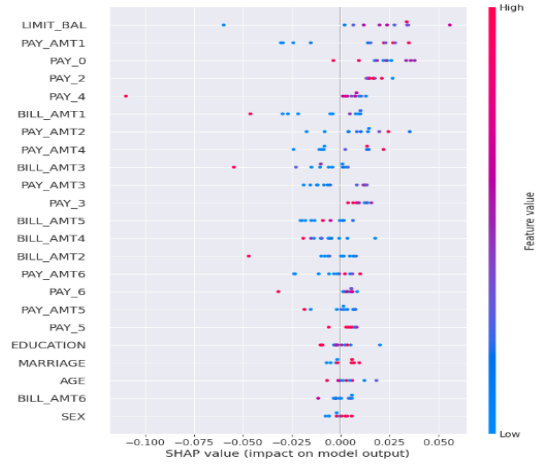

Figure 23 - Random Forest Classifier SHAP Results

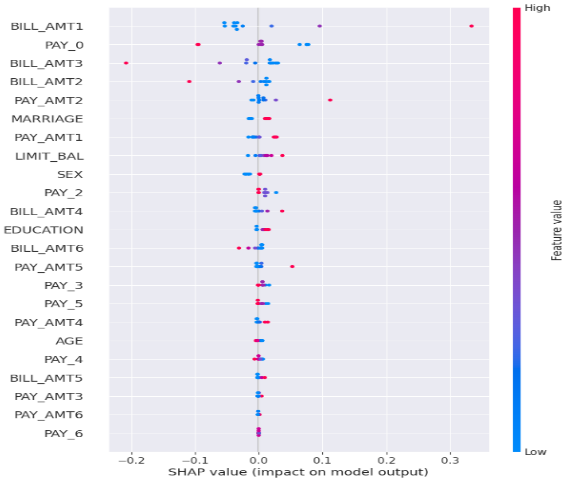

Figure 24 - Logistic Regression SHAP Results

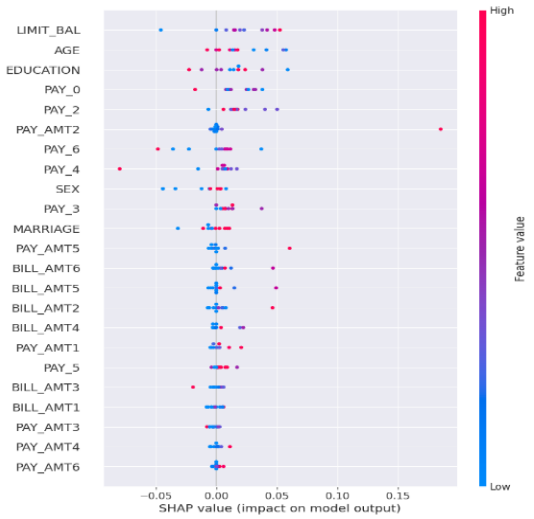

Figure 25 - K-Neighbors Classifier SHAP Results

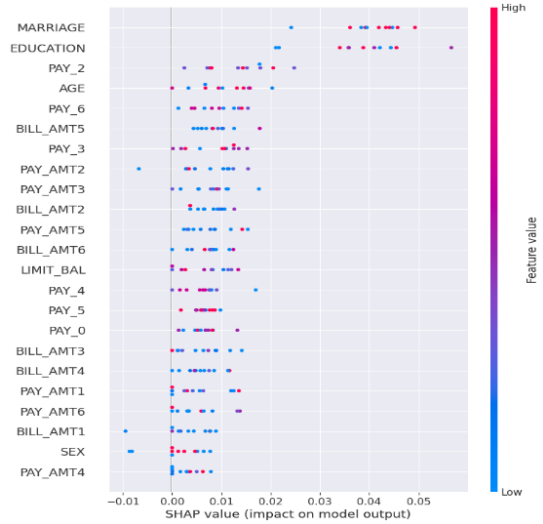

Figure 26 - Dummy Classifier SHAP Results

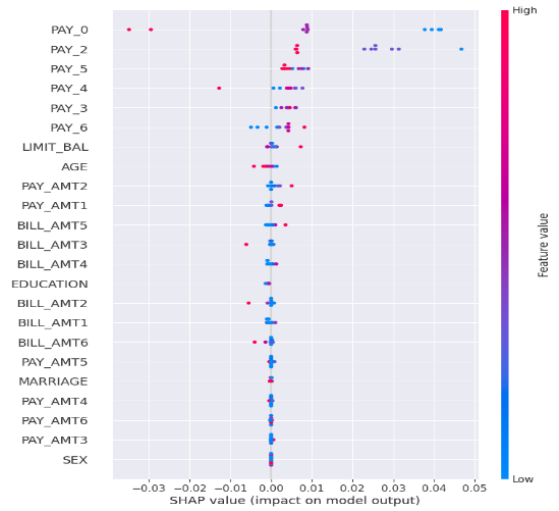

Figure 27 - SVC SHAP Results

We summarize the key findings of SHAP Results below:

- Payment History (PAY_X features): Across all models, the payment history appears to be a consistently strong predictor of the model's output, particularly the most recent payment status (PAY_0). Higher SHAP values indicate that recent payment behavior significantly influences the likelihood of a positive class prediction.
- Credit Limit (LIMIT_BAL): This feature has a prominent position in several models, with both positive and negative impacts on the model's prediction. It suggests that the credit limit is an important factor, but its effect can vary, possibly depending on the interaction with other features.
- Bill Amount (BILL_AMT_X features): The amount on the bill statements seems to have a variable impact on the models. For some models, the recent bill amounts (like BILL_AMT1) are more influential, while in others, the focus shifts to different statement periods

- Demographic Features (AGE, SEX, MARRIAGE, EDUCATION): These features appear throughout the plots but are less consistent in their positions and impact compared to payment history and credit limit. This suggests that while demographic information plays a role in the predictions, their influence may not be as strong or as direct as the financial history features.
- Repayment Amount (PAY_AMT_X features): Similar to bill amounts, the repayment amounts have varying impacts on the predictions. They do not stand out as the most influential factors but still contribute to the model's decisions.
- Consistency Across Models: While the features' influence varies across different models, there are patterns. Financial history (payment history, bill amounts, credit limit) consistently affects the predictions, whereas demographic factors show less consistent and generally weaker effects.
- Variability in Impact: The plots show a mix of positive and negative SHAP values for many features. This indicates that for some instances, a high feature value increases the likelihood of a positive prediction, while for others, it decreases it.

These summary plots indicate the relative importance of features across different machine learning models in a credit-defaulting context. Payment history is a key determinant in predictions, and other financial features also play significant roles, with demographic features contributing to a lesser and more varied extent. Moreover, the impact of a feature on the model output can be complex and depend on its interaction with other features.

e) **Bias Assessment:** In the following tables, we evaluate potential biases in credit default predictions within the mentioned machine learning models. For the purpose of analysis, the demographic attribute of sex and age are represented with binary indicators.

- *Sex=1 denotes Male*
- *Sex=0 denotes all other gender identities*
- *Age= 1 if Young Adulthood*
- *Age=0 denotes all other age groups*

Age Bias Key Findings (Table 8):

Bias Metric / ML Model	Accuracy			Equal Opportunity			Equalized Odds		
	Age=1	Age=0	Ratio	Age=1	Age=0	Ratio	Age=1	Age=0	Ratio
Random Forest Classifier	0.824	0.807	1.021	0.319	0.333	0.958	0.038	0.042	0.907
Logistic Regression	0.824	0.807	1.021	0.319	0.333	0.958	0.038	0.042	0.907
K-Neighbors Classifier	0.791	0.765	1.035	0.048	0.047	1.021	0.005	0.006	0.816
Dummy Classifier	0.657	0.647	1.017	0.208	0.209	0.996	0.22	0.214	1.026
SVC	0.824	0.807	1.021	0.319	0.333	0.958	0.038	0.042	0.907

Table 8 - Bias Results for Age

- Accuracy: The results show close accuracy ratios for Age=1 and Age=0 across models, which suggest similar predictive accuracy for both younger and older groups.

- Equal Opportunity: Ratios close to 1 indicate both age groups are similarly likely to be correctly predicted as credit-defaulting when they are actually default.
- Equalized Odds: Ratios close to 1 suggest that the models have similar true positive and false positive rates for both age groups.

Sex Bias Key Findings (Table 9):

Bias Metric / ML Model	Accuracy			Equal Opportunity			Equalized Odds		
	Sex=1	Sex=0	Ratio	Sex=1	Sex=0	Ratio	Sex=1	Sex=0	Ratio
Random Forest Classifier	0.801	0.831	0.965	0.336	0.314	1.07	0.048	0.034	1.422
Logistic Regression	0.801	0.831	0.965	0.336	0.314	1.07	0.048	0.034	1.422
K-Neighbors Classifier	0.763	0.797	0.957	0.059	0.038	1.533	0.008	0.004	2.399
Dummy Classifier	0.641	0.663	0.967	0.195	0.219	0.892	0.214	0.221	0.971
SVC	0.801	0.831	0.965	0.336	0.314	1.07	0.048	0.034	1.422

Table 9 - Bias Results for Sex

- Accuracy: Ratios near 1 suggest that models have similar accuracy for both sexes.
- Equal Opportunity: The K-Neighbors Classifier has a ratio significantly above 1, which may indicate a bias where one sex (possibly females if Sex=0) is less likely to be predicted to default compared to the other.
- Equalized Odds: The K-Neighbors Classifier and SVC display high ratios, indicating potential bias in how accurately they predict actual defaults and non-defaults between sexes.

In summary, for age, there's a slight bias indicated towards older individuals being predicted to default more than younger individuals. For sex, the K-Neighbors Classifier and SVC show potential bias in their predictions, favoring one sex over the other regarding the likelihood of defaulting. These observations would require further review and possibly model adjustments to address any unfair bias.

f) PrAIse Approach for Credit Default Scenario

For credit default prediction, the tool's analysis would reveal the most impactful features and the possible biases against older individuals and males, presuming a higher likelihood of default. These insights would emerge from the model explainability and bias evaluation phases, where the tool assesses how different features, such as age and sex, influence the risk assessment.

Upon identifying such biases, the tool would evaluate their risk level and provide recommendations. In the use case we covered, recommendations would be:

1. Examining and potentially expanding the dataset to include more diverse age groups and equal representation of sexes in various credit scenarios.
2. Implementing synthetic data generation methods to augment underrepresented groups in the training data.
3. Adjusting the model to incorporate age and sex as part of a fairness-aware learning framework that can recognize and correct for these biases.
4. Suggesting models that have a track record of less bias in demographic features when predicting creditworthiness.

4.5) Challenges and Limitations

The evaluation of explainable AI (XAI) techniques, specifically LIME and SHAP, has provided us with valuable insights into the decision-making of machine learning models, highlighting the ability to interpret predictions and identify embedded biases. However, our approach faced some challenges:

- LIME and SHAP have limitations, such as reliance on approximations that may not fully represent the true model workings, and difficulties scaling to large, high-dimensional datasets.
- Preprocessing techniques come with their own set of drawbacks. They can reduce interpretability, as the transformation or normalization of data might obscure the original feature meanings. Additionally, these techniques can lead to information loss that is critical for the model to learn from the data. The presence of outliers can also adversely affect these functions, making it crucial to identify and handle them appropriately during preprocessing to ensure accurate application of XAI methods.
- Certain model configurations aimed at improving explainability might require increased computational power and introduce greater complexity. This can impede a model's ability to find the optimal solution or to converge effectively, thus affecting performance and the quality of insights derived from the model.
- When considering feature effects, it is essential to be wary of proxy attributes or mutual information between features. These can subtly influence the impact of other features, leading to indirect biases or misconceptions about their importance. Furthermore, the

categorization of features can compromise the accuracy of insights. For example, binary classification of gender as "Male" and "non-Male" could oversimplify complex relationships within the data and contribute to biased outcomes or flawed interpretations.

The challenges outlined highlight the need for rigorous and mindful application of XAI techniques and preprocessing methods. Future research should focus on overcoming challenges such as establishing clear fairness metrics, enhancing model interpretability, and making bias-detection tools more accessible. Additionally, it is important for researchers to explore a variety of XAI tools beyond LIME and SHAP.

4.6) Summary

In the past, decisions regarding hiring, advertising, criminal punishment, and lending were typically made by humans and organizations, following federal, state, and local laws emphasizing fairness, transparency, and equality. However, today, machines have become significantly involved in these decisions, offering unparalleled efficiencies at scale and statistical precision [96]. Algorithms leverage vast macro- and micro-data to impact human behavior across different domains, from suggesting recruitment targets to aiding banks in assessing creditworthiness. However, as these services may seem time-efficient and productive, they may compromise transparency and equality in various industries. In scenarios beyond crediting and recruitment, automated risk assessments used by law enforcement agencies to determine bail and penalty limits can yield incorrect results, potentially resulting in significant cumulative effects on specific populations, such as imposing longer prison sentences or higher bail on individuals of color, or

even false convictions. Therefore, industries should fully comprehend the benefits and risks of implementing AI before doing so. This will be further discussed in the following chapter.

Our research highlights the role of explainability and bias detection tools in fostering fairness and accountability in AI development. And through the proposal of PrAIse, we show that providing developers with actionable insights into the presence of biases can help them take preemptive measures to mitigate potential harm and ensure equitable outcomes. Moreover, such tools can empower them to make informed decisions and uphold ethical standards throughout the model lifecycle. Overall, this work underscores the importance of incorporating ethical considerations into the workflow of smart IoT services, emphasizing the need for interdisciplinary collaboration between AI researchers, ethicists, and domain experts.

5. Decoding Justice: The Synergy of Artificial Intelligence and Machine Learning in the Legal Landscape

5.1) Introduction

The use of Artificial Intelligence (AI) and Machine Learning (ML) is progressing rapidly among most industries. As discussed in Chapter 2, given the diverse range of industries with distinct services and objectives, there is no universally applicable approach to govern AI implementation. Each environment is unique, requiring a nuanced approach. Before actual deployment, it is crucial to conduct comprehensive qualitative studies of AI implementation from a domain-specific and customized perspective. Crime prevention and effective resource allocation have long been top concerns for law enforcement agencies (LEAs) worldwide. By leveraging historical crime and sensory data and advanced analytical techniques, our study aims to contribute to the growing body of knowledge on effective crime prevention and resource allocation. The goal is to promote the transformation of law enforcement agencies from reactive responders into proactive crime preventers. This transformation is achieved through predictive analytics, enhanced surveillance with facial recognition, and efficient data analysis using natural language processing. Moreover, collaboration with AI experts ensures responsible deployment, balancing security needs with privacy concerns. Eventually, crime incidents can be detected, prevented, and solved at a much more accurate and faster rate, allocating more time for a wider range of offenses.

Serving as an extension of previous research, this chapter broadens the exploration of AI implementation beyond the realms of finance and employment, addressing the unique challenges and opportunities within law enforcement—a sector where the deeply moralistic and societal nature of law often lags behind technological advancements.

By reviewing literature on predictive policing, crime forecasting, and the integration of machine learning with security law enforcement, this chapter lays the groundwork for understanding AI's role in enhancing legal practices. It examines various AI functionalities and their applications within the legal framework, highlighting instances of authorities already employing AI to innovate law enforcement strategies. The potential risks associated with merging technology and law are also scrutinized, leading to a comprehensive discussion on how law enforcement agencies (LEAs) and AI/ML technologies can support and innovate secure, yet forward-thinking, security measures.

5.2) Background

AI and ML have been popularly introduced to the police, law enforcement, and crime reduction agencies to help discover where crime tends to be highest, aiding decision-making procedures that determine where to mark and install resources. This has been the focus of many researchers and academics, both in the technical and legal/ethical fields.

For Ogurlu *et al.*, the objective of their paper is to provide theoretical insights into the intersection of law and AI, highlighting potential challenges and solutions [98]. It also emphasizes the role of AI as a decision-making tool and focuses on the ongoing debate surrounding its current impact on the legal profession.

Chesterman *et al* discuss how the dual nature of integrating AI in legal practices presents a dilemma, as its utilization suggests a utilitarian perspective of law, where participants are seen merely as tools rather than individuals with inherent worth [99]. They believe that this prompts essential inquiries regarding the essence of law and authority: whether law can be simplified to algorithms aiming to maximize societal welfare, or if it must retain its essence as a domain of debate, politics, and association with institutions answerable to the public.

Similarly, Shah *et al* discuss how policymakers are forced to develop effective preventive measures to deal with the increasing criminal activities [100]. They study how the combination of ML and computer vision can be used by authorities and law agencies to identify, prevent, and reduce crimes at a faster rate to an evolutionary extent. They use the CCTVs in urban areas during a preliminary round to evaluate the initial productivity of such software, and then use supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning to perform the training process.

Belk focuses on the ethical dilemmas evident in five key areas - ubiquitous surveillance, social engineering, military robots, sex robots, and transhumanism - and how further advancements in technology will amplify concerns in these domains [101]. He explains how addressing these issues promptly through research is critical due to their profound implications.

According to Raaijmakers, LEAs law enforcement agencies are starting to rely on AI-enabled methods of data analysis and interpretation [102]. These applications include suspect profiling, traffic control, analyzing dark web money flows, and child abuse detection. The study discusses how researchers should develop tailored, flexible, and responsive solutions that can adapt to different contexts, roles, and end-user skills to effectively support law enforcement.

In Moses and Obi's research, the motivation was predicting automobile crimes and informing the responsible authorities to be prepared in advance for the crime occurrence [103]. Through this research, they describe how the predictive approach is a better solution to the automobile crime problem, and consequently encourage further research in this area.

According to Dakalbab *et al*, the growing shift toward technology, specifically advancements in AI, led to ML techniques being used for their ability to analyze large amounts of data in a shorter period and to extract crime patterns [104]. The authors also explain how analyzing crime records could reveal more information about the social structure of communities. Thus, decision makers and government entities will be able to better know which nationalities, age ranges, etc. to focus on in order to limit related problems.

In Ekanem's work, he considers how recent developments in Information Technology, analytics systems, and other mechanisms have been involving AI in crime control projects, and what positive impact this holds [105]. The work also considers the negative and positive contributions that this innovation will have in Nigeria. Therefore, the work proposes the need for governmental bodies to invest highly in modern crime-solving technologies, with AI, in Nigeria. Moreover, the research further recommends the need for implementing a law for regulating the use of modern technologies, specifically artificial intelligence in Nigeria.

Having examined the literature on the integration of AI and ML in law and its implications, benefitting or challenging, for law enforcement, the subsequent section will delve into specific use cases of AI/ML in law enforcement and policing, and will illustrate how the domains can work together through practical examples.

5.3) Use cases - AI and ML applications used for Smart Policing and Enforcement

With the digitized route the world is taking, AI and ML applications have and will continue to evolve in the field of law enforcement. Different types of AI and ML approaches exist, and they enable different uses and functionalities within law enforcement and policing. Some of the functionalities include:

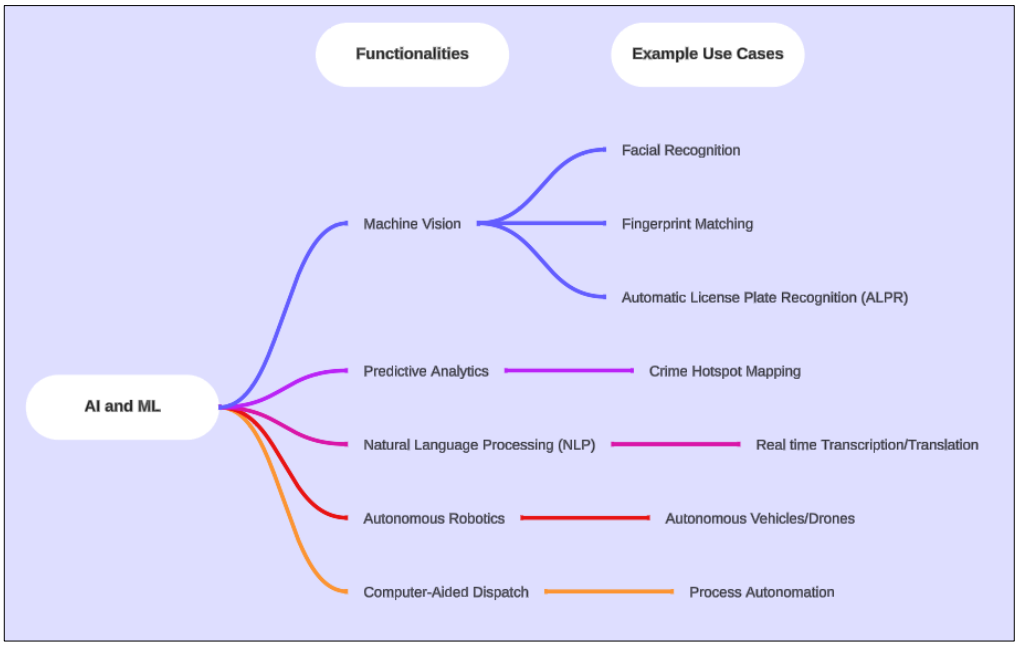

Figure 28 - Different types of AI/ML functionalities enable different applications within law enforcement.

a- Machine Vision: Also known as computer vision, machine vision involves the use of AI algorithms to analyze and interpret images or videos, allowing computers to understand their content and generate appropriate outputs [106]. Facilitated by ML, this process enables computers to classify, identify, verify, and detect objects within visual data. The common use cases of machine vision are facial recognition, fingerprint matching, and automatic license plate recognition (ALPR). Computer vision can be also employed in modern lie detector devices which analyze subtle facial and eye movements. One example is EyeDetect, a system that is used voluntarily by some employers in job interviews [106].

b- Predictive Analytics: AI is employed to enhance predictive policing models, both place-based and individual-based [107]. Place-based predictive policing analyzes data to pinpoint specific areas and times with heightened risk of crime or disorder. On the other hand, individual-based predictive policing utilizes information about individuals, such as their criminal justice history, to identify those at risk of involvement in future law enforcement incidents, whether as victims or offenders. A common application of predictive analysis is crime hotspot mapping and analysis with PredPol being one of the most widely used predictive policing program [108].

c- Natural Language Processing (NLP): NLP enhances the accuracy and efficiency of law enforcement reporting by allowing computers to understand, interpret, and generate human language in a way that is both meaningful and useful [107]. Common use cases include real time transcription and translation like automatic speech recognition (ASR) software. Rather than typing reports manually, officers are increasingly utilizing voice-to-text dictation tools, expediting the process. This technology not only accelerates transcription but also reduces human errors, such as misidentifying speakers' voices in witness testimonies and interrogations.

d- Autonomous Robotics: Autonomous robotic systems can be used by security services and law enforcement for surveillance and behavior detection [106]. They can also be of major use in areas of high danger risks and when dealing with objects of suspicion. For instance, to enhance the monitoring of public spaces, drones and self-driving vehicles equipped with AI-enabled sensors and cameras can be utilized. One example is the “Knightscope K5 Autonomous Data Machine”, a security patrol robot utilized in New York.

e- Computer-aided Dispatch (CAD): Law enforcement agencies utilize CAD systems to automate processes, gather call data, aid in resource allocation decisions, and facilitate reporting [107]. Providers like RapidDeploy are integrating AI and advanced analytics into CAD software,

offering benefits such as improved resource allocation, cost and time savings, enhanced predictions and decision making, and automated workflows.

After exploring multiple AI functionalities and their major possibilities for law and policy groups, we will delve into the actual endorsement and utilization of these AI applications.

5.4) Existing legal enforcement agencies that utilize AI ad ML

Several law enforcement agencies around the world have shown interest in and, in some cases, have endorsed the use of AI and ML technologies to enhance their capabilities. Here are a few examples:

- London Metropolitan Police (MET): The MET police utilized Live Facial Recognition (LFR) technology in Croydon, resulting in the arrest of over a dozen individuals during a two-day operation [109]. LFR cameras focused on a specific area streamed images of individuals passing through, cross-referencing them with a database containing a watchlist of wanted criminals, contributing to the controversial use of this technology in law enforcement for public safety.
- Singapore Police Force: The Singapore Police Force has introduced patrol robots at Changi Airport after more than five years of trials, with plans to deploy additional robots across the city-state in the future [110]. The robots, operational since April 2023, aim to enhance police presence, acting as "eyes on the ground" and assisting in incidents by enforcing cordons, warning bystanders, and facilitating direct communication with the public via a button on the robot's front, marking a significant advancement in the force's technological capabilities.

- Dubai Police: At GITEX Global Technology 2023, the Dubai Police introduced an AI-driven mechanism for handling minor traffic accidents [111]. This innovation analyzes minor accidents and autonomously generates reports for drivers, eliminating the need for human intervention. As part of the "Minor Accident Reporting" service on the Dubai Police app and website, this advancement showcases the integration of AI technologies to streamline accident reporting processes and enhance efficiency in law enforcement services.
- Chinese Authorities: In China, over 1.4 billion people are subject to constant surveillance through ubiquitous police cameras, phone tracking, purchase monitoring, and online chat censorship [112]. The Chinese police are investing in technology that utilizes extensive surveillance data to predict potential crime and protests in advance. These systems focus on individuals deemed suspicious by algorithms and Chinese authorities, irrespective of any wrongdoing on their part, reflecting a widespread use of advanced surveillance technologies for preemptive law enforcement measures.
- Los Angeles Police Department (LAPD): Machine learning will be employed to analyze the language used by LAPD officers during traffic stops, aiming to assess whether police language contributes to unnecessary escalations in public encounters [113]. The findings will inform officer training, enhancing their ability to navigate interactions with the public appropriately and promoting accountability, marking a significant advancement in the use of machine learning for law enforcement training.

Incorporating AI and ML into law enforcement and policing strategies can enhance protective measures across communities. However, despite the major safety accomplishments such

advanced technology measures can produce, governments deploying AI and ML must prioritize transparency, safety, and fairness as various risks are associated with such advancements.

5.5) Risks Associated with Merging AI and Law Enforcement

AI and ML systems can lead to bad outcomes as much as they can lead to positive ones if legal or governmental groups deviate from keeping approaches human-centered.

Some of the major risks associated with integrating AI and ML into law enforcement strategies include:

- Bias and Discrimination: AI implementation extends beyond work aspects to impact human lifestyles. Algorithms and training data, originating from human input, may inadvertently perpetuate harmful biases. This can result in AI systems generating unfair outcomes and fostering discrimination against minority groups who didn't commit any offense [114]. For instance, PredPol is an algorithmic predictive policing system that forecasts when and where property crimes are likely to occur. However, researchers found that PredPol may be creating a feedback loop, causing officers to patrol specific neighborhoods—often those with high racial minority populations—regardless of the actual crime rates in those areas [115].
- Misuse and Data breach: Cybercriminals can exploit AI capabilities for malicious purposes, developing machine learning models that manipulate inputs to facilitate cyberattacks. In March 2021, a bold cyber-attack compromised over 150,000 video surveillance cameras across various US organizations' cloud networks [116]. This poses a threat to individuals' privacy and freedom as AI systems, with their extensive data processing abilities, may store personal data for extended periods. The potential misuse

of collected data without consent raises concerns about privacy violations. Additionally, unintended data collection from individuals not initially targeted can compromise data integrity.

- Mishaps and Liability Concerns: The autonomous nature of AI systems enables them to generate novel decisions and data independently. However, the foreseeability requirement in negligent liability becomes challenging when autonomous systems exhibit unforeseen or socially wrongful behaviors, raising complex questions about liability attribution. This unpredictability may result in unforeseeable harms or drawbacks not anticipated by developers, sparking ethical controversies. In 2020, Robert Williams was wrongfully convicted and arrested for allegedly stealing thousands of dollars of watches because the Detroit police had matched grainy surveillance footage to his license photo using facial recognition [117]. While some argue that human developers and operators bear a duty of care for the decisions made by the AI systems they create, others propose assigning legal personalities to autonomous machines to give them liability, which in turn would lead to extended ethical challenges and debates.

It is clear that advancements in societal and legal aspects constitute a double-edged sword. By effectively managing the associated risks, groups can fully enjoy the benefits and advantages of these innovations.

5.6) Discussion

Incorporating AI into crime prevention and mitigation, under human supervision, holds promising results for both the Technology and Law domains. This integration carries mutual benefit to both law enforcement agencies and AI developers and analysts.

According to Deloitte, the implementation of smart technology, such as AI and ML, could potentially result in a 30 to 40% reduction in crime rates within cities [118]. ML methods can be used to identify obscure patterns and irregularities that might go unnoticed by human analysts. While one of the risks associated with technological advancements is bias, AI and ML can help reduce bias by detecting discriminatory behaviors and information within systems or datasets. This in turn leads to uncovering hidden networks and connections contained within not only criminal events but also social issues and imperfections. Moreover, ML models can analyze real-time data streams, such as social media inputs or surveillance camera footage, and that provides law enforcement agencies with direct insights into ongoing criminal activities. This consequently helps law enforcement agencies to take proactive measures, perform predictive policing, and prevent crimes by allocating resources, personnel, and assets, effectively and deterring criminal activities before they occur. Furthermore, automating certain tasks within the context of law enforcement not only liberates time and resources but also enables a more focused approach to addressing crime-related issues. The time saved can be channeled towards dealing with non-automatable crime concerns and implementing preventive measures, thereby contributing to a more comprehensive crime prevention strategy. This dual impact of automation, both in proactive prevention and responsive management, underscores its significance in optimizing law enforcement efforts and resource utilization.

In parallel, LEAs play a role in helping machine learning developers and architects achieve promising results in smart crime prevention. They generate a wide range of crime-relevant data – incidents, social dynamics, and use cases. Developers and analysts can then use this data to train and improve the accuracy of machine learning models for several types: patten recognition, anomaly detection, classification, clustering, and real-time analysis. By providing real-life

scenarios, the feedback loop from deployment to model refinement improves the efficiency of the technology. Moreover, the merge of law enforcement and machine learning can help further develop ethical standards and guidelines for safe technology tackling privacy concerns, biases, and misuse and abuse. Additionally, when the public perceives that AI practices are endorsed and supported by the law, it fosters a greater receptivity and acceptance among the populace. This endorsement not only bolsters public confidence but also encourages an embrace of AI technologies. The assurance of legal support provides a foundation for trust, paving the way for the integration of AI practices into societal norms.

What is critically essential to comprehend is that to the extent that AI aids law enforcement, law can reciprocally support AI. According to a study conducted by the Pew Research Center, 45% of adults express both concern and enthusiasm regarding the increasing utilization of AI in law enforcement [118]. Ensuring transparent, fair, and consented goals, particularly oriented toward human advantages, sustains the progression of both technical and legal domains. Consequently, it is paramount for AI and ML developers to ensure accuracy, lack of bias, and precision in their designs. This responsibility aligns with that of legal authorities, who must stay abreast of advanced technology trends, learn effective governance, and integrate these technologies into safety and crime prevention practices for the greater good. To accomplish this successful Tech-Law vision, here are a few recommendations:

- Achieving a seamless integration and merger of technology and law necessitates a thorough comprehension of all relevant terminology, definitions, and results within both fields. To foster reasoning and transparency of AI techniques and outcomes, legal agencies can use Explainable AI (XAI) tools that explain the decision-making process of AI and how and why certain predictions were made. As for the comprehension of legal

notions, analysts can depend on Legal Information Institutes (LIIs), which are location-specific, and provide free access to databases of case law, statutes, regulations, and other legal materials, along with explanations and summaries of legal concepts [119].

- To ensure the responsible use of AI in LEA systems, it is imperative to establish clear regulatory measures. This was thoroughly mentioned in the previous chapter too. Manufacturing firms and adopting institutions should adhere to guidelines, codes of practice, conditions, and principles and should always align with international human rights. This ensures that the design and implementation of these advanced systems are not exploited for profit at the expense of human rights or manipulated by staff to evade significant responsibilities [114]. Examples of existing AI Governance frameworks include: The National Institute of Standards and Technology (NIST), and The Model AI Governance Framework (PDPC), the Implementation and Self-Assessment Guide for Organizations (ISAGO), and recently, the Artificial Intelligence and Data Act (AIDA), and the EU AI Act [69].

- "For government applications - particularly those with high costs to a false positive - it is important to have a human in the loop… Preserving the final decision with an individual not only helps pattern recognition technology be more accurate, but it also clarifies accountability when mistakes do occur." said Devora Kaye, an NYPD spokesperson [120]. Pattern recognition tools and predictive modeling can hold favorable results when joined by human analysts. Whether it’s for algorithmic auditing, testing and verification before implementation, or distant supervision, maintaining a human-in-the-loop system

allows humans to stay in control, minimizing unforeseeable damage and preserving accountability and liability during unfortunate occurrences.

- With intentions to promote fairness and equity of AI techniques, especially in applications like predictive policing, facial recognition, and risk assessment, legal agencies can adopt tools like Google's "Fairness Indicators" or IBM's "AI Fairness 360" to evaluate the fairness of AI and ML models and detect and mitigate bias in data and algorithms. Also, governmental groups and analysts should focus on viewing AI models as one part rather than a complete accurate investigator. Wrongfully convicted Robert Williams believes that it should be treated "as an investigative lead, not as the only proof they need to charge someone with a crime" [117].

- Multi-Stakeholder Collaboration: Law enforcement agencies should collaborate with academic institutions, civil society organizations, and technology experts to develop and implement best practices for the ethical and safe use of AI. For example, the Police Data Initiative in the United States fosters collaboration between police departments and data scientists to enhance transparency and accountability in policing through data-driven approaches [121]. Also, ethicists should carry on with further debates and discussions regarding the social and ethical challenges surrounding the advancements, to help clarify the ambiguity regarding current ones and pave the way for the new ones.

5.7) Summary

The delicate nature of technology allows for both positive and negative applications. Yet, when directed with human-focused approach, incorporating cutting-edge technologies into legal systems can address and diminish societal issues yielding numerous benefits. Respectively, the proactive adoption of ML and AI by LEAs serves as a compelling demonstration of the positive and trustworthy impact of advanced technology on societies. By embracing these advancements, LEAs not only bolster their operational capabilities but also contribute to the broader narrative of utilizing technology for the betterment of communities, thereby reinforcing the symbiotic relationship between technological innovation and societal welfare.

This chapter addresses an unexplored niche within the intersection of law and AI by delving into its complexities, offering comprehensive insights into existing applications, endorsements, successes, and challenges. By examining both sides of this dual-edged sword, this work not only provides practical recommendations but also advocates for a safer, more just, and transparent integration of AI in legal practices. Another contribution of this chapter would be to address how all domains and industries should navigate the inevitable growth of AI. Therefore, this work can serve as a blueprint for domains beyond law, providing them with insights into considerations and approaches when implementing AI. This fills a crucial gap in literature and encourages progress towards a more equitable and ethically sound innovative future.

6. Conclusion and Future Work

6.1) Conclusion

Through this research, our aim is to contribute to the establishment and preservation of human-centered standards in the realm of AI advancements and smart systems. These standards encompass human involvement in decision-making, clarity regarding AI deployment and understanding of system procedures, accuracy and reliability of datasets, transparent communication between staff and stakeholders, alignment of outcomes with initial objectives, and effective capability for risk and bias mitigation. If these standards were to be given serious attention by tech workers and developers as a generic policy-making framework, then general and common risks associated with smart advancements and AI would diminish. Regarding eXplainable AI (XAI) tools, we examine how they help ensure transparency, fairness, and clarity throughout the processes of machine learning predictions. This transparency empowers decision-makers to make more informed and reasonable choices. Additionally, we explore common methods used to identify potential biases or inaccuracies in the model with the hopes of reducing them afterward. Despite their utility, the growing complexity of AI algorithms poses a growing difficulty in comprehending and deciphering how and why a specific outcome was reached. Therefore, we also discuss current and possible challenges and limitations that can hinder the promised role of XAI and consequently exacerbate unwanted bias. The hope and motivation are that this work and its results will inspire researchers and stakeholders to expand our approach and navigate the general and focused impact of AI in their respective fields, fostering continued improvement and fairness in the development and deployment of AI technologies and advancements.

As AI applications are increasingly deployed in contexts where they wield considerable influence over life-altering decisions, ensuring fairness and transparency becomes paramount. In summary, the contributions of the presented chapters are the following:

The second chapter delves into the intricate relationship between AI and IoT systems, offering insights into their definitions, historical evolution, and methodological considerations. By analyzing their union, the chapter provides a comprehensive understanding of AI-enabled IoT systems, along with recommendations for optimizing their potential while addressing associated challenges.

The third chapter underscores the nuanced governance required for advanced technical systems like AI and IoT, emphasizing the absence of a one-size-fits-all approach. It elucidates the evolution of governance frameworks from design to implementation phases and explored strategies for fostering transparency and trust in AI, crucial for ensuring ethical practice and effective communication.

In the fourth chapter, actionable insights are provided for developers to detect and mitigate biases during the design phases of AI systems. This contribution aids in promoting fairness and equity in AI decision-making processes.

The fifth chapter extends the discussion beyond law enforcement, serving as a blueprint for various domains seeking to implement AI. By offering considerations and qualitative approaches, it empowers domains to navigate the complexities of AI deployment effectively.

6.2) Future Work

Looking towards future directions, researchers are encouraged to address multiple considerations:

- Exploring new areas: Research should address additional biases, and expand AI administration analysis into diverse domains beyond law enforcement. Moreover, a thorough exploration of recommendations beyond transparency and fairness is warranted, recognizing that every aspect contributes to safer and more innovative AI deployment across all domains.
- Distinguishing Between AI's Potential and Fiction: Research should aim to clearly delineate practical applications of AI from speculative fiction, fostering a balance between visionary ideas and their pragmatic execution. This involves demystifying AI to dispel unfounded fears and overenthusiasm, using these perceptions constructively to develop a systematic framework for AI’s future.
- Safety and Standards in AI Integration: Investigating the safety implications of incorporating AI into IoT systems compared to traditional methods is crucial. Research should prioritize the physical, mental, and digital well-being of individuals, setting clear standards for both traditional and advanced systems to guide decision-making and improve risk areas.
- Data Design and Security: As data management becomes increasingly complex, future studies must explore innovative strategies for data design, collection, storage, and security within IoT architectures. This research should ensure compatibility with existing communication standards and address the specific needs of organizations and customers.
- Operational and Management Efficiency: Further research is needed to optimize the operational workflow of AI/IoT systems. This includes the development of advanced data analytics tools, enhancing data loss compensation mechanisms, and establishing trust checkpoints for data sources to improve system efficiency and reliability.

- Governance and Communication Strategies: The implementation of AI/IoT requires effective governance structures and clear communication with all stakeholders involved. Future research should focus on the formation of governing committees, defining roles and responsibilities, and developing comprehensive communication strategies to inform stakeholders about AI impacts, risks, and feedback mechanisms.
- Legal Reasoning and Transparency with XAI: Exploring how legal agencies can utilize XAI tools to make AI decision-making processes more transparent and understandable is a vital research avenue. This includes analyzing legal terms, outcomes, and the application of XAI in interpreting legal documents and decisions.
- Compliance with Ethical and Legal Standards: Lastly, achieving seamless integration of AI technologies into society necessitates ongoing research into compliance with ethical and legal standards. This encompasses understanding the intersection of technology and law, fostering transparency, and ensuring AI applications are ethical and legally sound.

Bibliography

[1] Woetzel, J et al (2018, June 5). Smart cities: Digital solutions for a more livable future. McKinsey Global Institute. Retrieved from https://www.mckinsey.com/business-functions/operations/our-insights/smart cities-digital-solutions-for-a-more-livable-future

[2] Vinugayathri. (2020).AI and IoT Blended - What It Is and Why It Matters? Retrieved from https://www.clariontech.com/blog/ai-and-iot-blended-what-it-is-and-why-it-matters

[3] PwC (2017). Leveraging the upcoming disruptions from AI and IoT. Retrieved from https://www.pwc.com/gx/en/industries/communications/assets/pwc-ai-and-iot.pdf

[4] Hashimoto, D and others. (2018). "Artificial Intelligence in Surgery: Promises and Perils." Annals of surgery, 268(1), 70-76

[5] Perez, j and others (2018). Artificial Intelligence and Robotics. UK-RAS Network

[6] Middleton. M (2021). Deep Learning vs Machine Learning — What's the Difference?. Flatiron School. Retrieved from https://flatironschool.com/blog/deep-learning-vs-machine-learning

[7] IBM cloud education (2020). Strong AI. Retrieved from https://www.ibm.com/cloud/learn/strong-ai

[8] Sharma, N and others (2019). The History, Present and Future with IoT. Internet of Things and Big Data Analytics for Smart Generation, Intelligent Systems Reference Library 154. Springer Nature Switzerland AG

[9] Lueth, K (2014). Why the Internet of Things is called Internet of Things: Definition, history, disambiguation. Retrieved from https://iot-analytics.com/internet-of-things-definition/

[10] Thaler, D and others (2015). Architectural Considerations in Smart Object Networking. IAB

RFC 7452. Retrieved from https://www.ietf.org/proceedings/92/slides/slides-92-iab-techplenary-2.pdf

[11] Swamy, N and others (2016). Analysis on IoT Challenges, Opportunities, Applications and Communication Models. International Journal of Advanced Engineering, Management and Science. 2(4)

[12] Shah, R and Chircu, A (2018). IOT AND AI IN HEALTHCARE: A SYSTEMATIC LITERATURE REVIEW. Issues in Information Systems 19(3), pp. 33-41.

[13] Atreyi Kankanhalli, A and others (2019). IoT and AI for Smart Government: A Research Agenda. Government Information Quarterly, 36(2), 304-309.

[14] Huertas-Lopez, C et al (2021). Artificial Intelligence: Transformation Of The Roles Of Human Capital. International Journal of Scientific & Technology Research. 10 (9).

[15] N. Kumar, N. Kharkwal, R. Kohli and S. Choudhary (2016). "Ethical aspects and future of artificial intelligence; 2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH), pp. 111-114.

[16] Puntoni, S., Reczek, R. W., Giesler, M., & Botti, S. (2021). Consumers and Artificial Intelligence: An Experiential Perspective. Journal of Marketing, 85(1), 131–151. https://doi.org/10.1177/0022242920953847

[17] Banafa, A. (2019). Secure and Smart Internet of Things (IoT): Using Blockchain and Artificial Intelligence (AI). Stylus Publishing, LLC.21

[18] Kumar, S (2019). Advantages and Disadvantages of Artificial Intelligence. Retrieved from https://towardsdatascience.com/advantages-and-disadvantages-of-artificial-intelligence-182a5ef6588c

[19] Amato, F., López, A., Peña-Méndez, E. M., Vaňhara, P., Hampl, A., & Havel, J. (2013).

Artificial neural networks in medical diagnosis. Journal of Applied Biomedicine (De Gruyter Open), 11(2), 45-58.

[20] Kaur, M.J., Mishra, V.P., Maheshwari, P. (2020). The Convergence of Digital Twin, IoT, and Machine Learning: Transforming Data into Action. In: Farsi, M., Daneshkhah, A., Hosseinian-Far, A., Jahankhani, H. (eds) Digital Twin Technologies and Smart Cities. Internet of Things. Springer, Cham. https://doi.org/10.1007/978-3-030-18732-3_1

[21] Burns, Larry. 2017. Autonomy: The New Age of Automobility, Autonomous Vehicle Engineering.

[22] Ashrafian, H. (2015). AIonAI: A Humanitarian Law of Artificial Intelligence and Robotics Science and Engineering Ethics, 21, 29-40.

[23] Upchurch, M. (2018). Robots and AI at work: the prospects for singularity. New Technology, Work and Employment, 33(3), 205-218.

[24] Osullivan, S et al. (2018). Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery;. The International Journal of Medical Robotics and Computer-Assisted Surgery, 15(1).

[25] Dilmegani, C. (2021, November 11) “AI Ethics: Top 9 Ethical Dilemmas of AI and How to Navigate Them”. Retrieved from https://research.aimultiple.com/ai-ethics/

[26] Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L (2016). The ethics of algorithms: Mapping the debate. Big Data & Society.

[27] Belani, G. (2021) “The Use of Artificial Intelligence in Cybersecurity: A Review”. Retrieved from https://www.computer.org/publications/tech-news/trends/the-use-of-artificial-intelligence-in-cybersecurity22

[28] Pearce, G. (2021, May 28) Beware the Privacy Violations in Artificial Intelligence Applications. Retrieved from https://www.isaca.org/resources/news-and-trends/isaca-now-

blog/2021/beware-the-privacy-violations-in-artificial-intelligence-applications

[29] Hoffman, D. L. and Thomas P Novak (2018), Consumer and Object Experience in the Internet of Things: An Assemblage Theory Approach, Journal of Consumer Research. 44(6), 1178–1204.

[30] Laitinen, A and Sahlgren, O. (2021, October 26) “AI Systems and Respect for Human Autonomy”. Front. Artif. Intell. Retrieved from https://doi.org/10.3389/frai.2021.705164

[31] Bossmann, J. (2016, October 21) “Top 9 ethical issues in artificial intelligence”. Retrieved from https://www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/

[32] Aimun A.B. Jamjoom and others, ‘Exploring public opinion about liability and responsibility in surgical robotics with the iRobotSurgeon survey’, [2020]

[33] Nick hilborne, Let robots own property, Supreme Court justice suggests (Legal Futures, 19 March 2019) https://www.legalfutures.co.uk/latest-news/let-robots-own-property-supreme-court-justice-suggests; accessed 1 August 2021

[34] David S. Kemp, D (2012). 'Autonomous Cars and Surgical Robots: A Discussion of Ethical and Legal Responsibility. Retrieved from https://verdict.justia.com/2012/11/19/autonomous- cars-and-surgical-robots

[35] Prince, C. (2018) Duty of Care Can Robots Replace Humans? Just Ask Elon Musk. SHRM. Retrieved from https://www.shrm.org/resourcesandtools/hr-

[36] Castellanos, S. (2020). “AUTONOMOUS ROBOTS ARE COMING TO THE OPERATING ROOM”. The Wall Street Journal.

[37] Svoboda, E (2019). ‘Your robot surgeon will see you now’. Retrieved from https://www.nature.com/articles/d41586-019-02874-0

[38] Rozbruch, L (2018, March 16). Litigation & Robotic Surgery: Product Liability or Medical Malpractice? The Roundtable. Retrieved from https://www.pulj.org/the-roundtable/litigation-robotic-surgery-product liability-or-medical-malpractice

[39] Bogue, R (2014). ‘Robot ethics and law Part two: law. Industrial Robot: An International Journal, 41(5), 398-402

[40] PDPC. (2020). Singapore’s Approach to AI Governance. Retrieved from https://www.pdpc.gov.sg/Help-and-Resources/2020/01/Model-AI-Governance-Framework.

[41] Usluoğulları, F and others. (2017). Robotic surgery and malpractice. Turkish Journal of Urology, 43(4), 425-428.

[42] V. A. F. Almeida, D. Doneda and M. Monteiro: Governance Challenges for the Internet of Things. IEEE Internet Computing. Vol. 19, no. 4, pp.

[43] Kulkarni, K.: Smart City as System of Systems: Subject of study - Vertical Farming and Autonomous Driving in Smart city. INCOSE International Symposium, 29: 505-517. (2019)

[44] N. V. Lopes,: Smart governance: A key factor for smart cities implementation. IEEE International Conference on Smart Grid and Smart Cities (ICSGSC). https://ieeexplore.ieee.org/abstract/document/8038591 (2017)

[45] Joshi, S et al.: Developing Smart Cities: An Integrated Framework. https://www.sciencedirect.com/science/article/pii/S1877050916315022 (2016)

[46] Dameri, R. P. and Benevolo, C.: Governing Smart Cities: An Empirical Analysis. Social Science Computer Review, 34(6), 693–707. (2016)

[47] Yigitcanlar, T et al.: Understanding smart cities: Intertwining development drivers with desired outcomes in a multidimensional framework. https://www.sciencedirect.com/science/article/abs/pii/S0264275117313367 (2018)

[48] Kroger, W: Automated vehicle driving: background and deduction of governance needs, Journal of Risk Research, 24:1, 14-27. (2017)

[49] Bundin, M et al.: Legal framework for self-driving cars: the case of Rus- sia. In Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance. Association for Computing Machinery, New York, NY, USA, 206–213. (2020)

[50] Cohen, T et al.: Reframing the governance of automotive automation: insights from UK stakeholder workshops, Journal of Responsible Innovation, 5:3, 257-279, (2018)

[51] M. J. Hanna and S. C. Kimmel.: Current US Federal Policy Framework for Self-Driving Vehicles: Opportunities and Challenges, in Computer, vol. 50, no. 12, pp. 32-40, (2017)

[52] Wolfert, S et al.: Big Data in Smart Farming – A review. https://www.sciencedirect.com/science/article/pii/S0308521X16303754 (2017) Accessed July 2022 [53] D.Boursianis, A et al.: Internet of Things (IoT) and Agricultural Unmanned Aerial Vehicles (UAVs) in smart farming: A comprehensive review. https://www.sciencedirect.com/science/article/abs/pii/S2542660520300238 (2020) Accessed July 2022

[54] Debauche, O et al.: Data management and internet of things: A methodological review in smart farming. https://www.sciencedirect.com/science/article/abs/pii/S2542660521000226 (2021)

[55] IMDA: Internet of Things (IoT) Cyber Security Guide. https://www.imda.gov.sg/-/media/Imda/Files/Regulation-Licensing-andConsultations/ICT-Standards/Telecommunication-Standards/ReferenceSpec/IMDA-IoT-Cyber-Security-Guide.pdf (2020) Accessed July 2022

[56] Gerard, M.: Standards for Cybersecure IoT Devices: A Way Forward. https://www.cigionline.org/publications/standards-cybersecure-iotdevices-way-forward/ (2020) Accessed July 2022

[57] CSA IoT Security Controls Framework. Retrieved from https://cloudsecurityalliance.org/artifacts/iot-security-controlsframework/ (2019)

[58] Infocomm Media Development Authority (IMDA) and Personal Data Protection Commission (PDPC). A Guide to Job Redesign in the Age of AI. https://file.go.gov.sg/ai-guide-to-jobredesign.pdf (2020). Accessed July 2022

[59] Companion to the Model AI Governance Framework – Implementation and Self-Assessment Guide for Organizations. https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGIsago.pdf (2020) Accessed June 2022

[60] ISACA: ISACA Shares Good IoT Governance for Safe and Efficient Deployment. https://www.isaca.org/why-isaca/about- us/newsroom/pressreleases/2017/isaca-shares-good-iot-governance-for- safe-and-efficientdeployment (2017) Accessed July 2022

[61] What is explainable AI. IBM. Retrieved from https://www.ibm.com/topics/explainable-ai

[62] Gramegna A and Giudici P (2021) SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk. Front. Artif. Intell. 4:752558. doi: 10.3389/frai.2021.752558

[63] D. Nedeljkovic, N. Y. Fares, and M. Jammal, "A Novel Machine Learning-based Approach to City Crime Sensor Placement Prediction," Accepted at IEEE 9th World Forum on Internet of Things, 2023.

[64] Rueda-Rueda et al. Framework-based security measures for Internet of Thing: A literature review Open Computer Science, vol. 11, no. 1, 2021, pp. 346-354. https://doi.org/10.1515/comp-2020-0220

[65] Keita, Z, "Explainable AI - Understanding and Trusting Machine Learning Models". Retreived from https://www.datacamp.com/tutorial/explainable-ai-understandingand-trusting-machine-learning-models

[66] Model Artificial Intelligence Governance Framework and Assessment Guide. https://www.weforum.org/projects/model-ai-governance- framework (2022). Accessed July

[67] Boukherouaa, E. B., Shabsigh, M. G., AlAjmi, K., Deodoro, J., Farias, A., Iskender, E. S., ... & Ravikumar, R. (2021). *Powering the digital economy: opportunities and risks of artificial intelligence in finance*. International Monetary Fund.

[68] Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., ... & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artificial Intelligence, 296, 103473.

[69] N. Y. Fares, D. Nedeljkovic, M. Jammal, "AI-enabled IoT Applications: Towards a Transparent Governance Framework" Accepted at IEEE 7th Global Conference on Artificial Intelligence and Internet of Things, 2023

[70] Casey, B., Farhangi, A., & Vogl, R. (2019). Rethinking explainable machines. Berkeley Technology Law Journal, 34(1), 143-188.

[71] Praveenraj, D. D. W., Victor, M., Vennila, C., Alawadi, A. H., Diyora, P., Vasudevan, N., & Avudaiappan, T. (2023). Exploring Explainable Artificial Intelligence for Transparent Decision Making. In E3S Web of Conferences (Vol. 399, p. 04030). EDP Sciences.

[72] Finkel, N. J. (2000). But it's not fair! Commonsense notions of unfairness. Psychology, Public Policy, and Law, 6(4), 898.

[73] Merriam-Webster. (n.d.). *Fairness*. Retrieved from https://www.merriam-webster.com/dictionary/fairness

[74] Cambridge Dictionary. (n.d.). Fairness. Retrieved from https://dictionary.cambridge.org/dictionary/english/fairness

[75] Smith, G., Kohli, N., & Rustagi, I. (2020). What does "fairness" mean for machine learning systems. Berkeley Haas EGAL.

[77] Reagan, M (2021). Understanding Bias and Fairness in AI Systems. Retrieved from https://towardsdatascience.com/understanding-bias-and-fairness-in-ai-systems-6f7fbfe267f3

[78] Bacelar, M. (2021). Monitoring bias and fairness in machine learning models: A review. *ScienceOpen Preprints*.

[79] Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020, February). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 180-186).

[80] Dieber, J., & Kirrane, S. (2020). Why model why? Assessing the strengths and limitations of LIME. arXiv preprint arXiv:2012.00093.

[81] Arias-Duart, A., Parés, F., Garcia-Gasulla, D., & Gimenez-Abalos, V. (2022, July). Focus! rating XAI methods and finding biases. In 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp. 1-8). IEEE.

[82] Bertrand, A., Belloum, R., Eagan, J. R., & Maxwell, W. (2022, July). How cognitive biases affect XAI-assisted decision-making: A systematic review. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (pp. 78-91).

[83] Suffian, M., & Bogliolo, A. (2022). Investigation and Mitigation of Bias in Explainable AI. In CEUR WORKSHOP PROCEEDINGS (Vol. 3319, pp. 89-94).

[84] UN. Universal Declaration of Human Rights. Retrieved from https://www.un.org/en/about-us/universal-declaration-of-human-rights#:~:text=Article%2023,equal%20pay%20for%20equal%20work.

[85] ILO. R111 - Discrimination (Employment and Occupation) Recommendation, 1958 (No. 111). Retrieved from https://www.ilo.org/dyn/normlex/en/f?p=NORMLEXPUB:12100:0::NO::P12100_INSTRUMENT_ID:312449

[86] Birkelund, G. E., Lancee, B., Larsen, E. N., Polavieja, J. G., Radl, J., & Yemane, R. (2022). Gender discrimination in hiring: Evidence from a cross-national harmonized field experiment. European Sociological Review, 38(3), 337-354.

[87] Palmer, A (2023). Amazon sued by three employees who allege gender discrimination and 'chronic' pay inequity. Retrieved from https://www.cnbc.com/2023/11/20/amazon-sued-by-three-employees-who-allege-gender-discrimination.html#:~:text=Amazon%20has%20faced%20allegations%20of,in%20its%20cloud%20computing%20unit.

[88] Quillian, L., Heath, A., Pager, D., Midtbøen, A. H., Fleischmann, F., & Hexel, O. (2019). Do some countries discriminate more than others? Evidence from 97 field experiments of racial discrimination in hiring. Sociological Science, 6, 467-496.

[89] Gleeson, C (2024). Why and How to Use AI and Automation for Recruitment? Retrieved from https://www.occupop.com/blog/why-and-how-to-use-ai-and-automation-in-recruitment

[90] HR Analytics Job Change of Data Scientists. Retrieved from https://www.kaggle.com/datasets/arashnic/hr-analytics-job-change-of-data-scientists

[91] Fortney, J (2021). Recruitment Targets ML. Retrieved fromhttps://github.com/JordanFortney/recruitmentTargetsML

[92] The World Bank Group (2019). CREDIT SCORING

APPROACHES GUIDELINES. Retrieved from https://thedocs.worldbank.org/en/doc/935891585869698451-0130022020/original/CREDITSCORINGAPPROACHESGUIDELINESFINALWEB.pdf

[93] Office of Public Affairs (2024). Justice Department Secures Agreement with Patriot Bank to Resolve Lending Discrimination Claims. Retrieved from https://www.justice.gov/opa/pr/justice-department-secures-agreement-patriot-bank-resolve-lending-discrimination-claims

[93] The New York Times (2019). Retrieved from https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html

[94] Marques, S., Mariano, J., Mendonça, J., De Tavernier, W., Hess, M., Naegele, L., ... & Martins, D. (2020). Determinants of ageism against older adults: A systematic review. International journal of environmental research and public health, 17(7), 2560.

[95] Wang, K (2023). Credit default Prediction Group. Retrieved from https://github.com/UBC-MDS/credit_default_prediction_group_20?tab=readme-ov-file

[96] Default of Credit Card Clients. Retrieved from https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients

[97] Lee, E. (2021). How do we build trust in machine learning models?. Available at SSRN 3822437.

[98] Ogurlu, Y., Shagieva, R. V., & Bersanov, A. S. (2021, April). A debate on artificial intelligence in area of law and the legal professions. In Institute of Scientific Communications Conference (pp. 365-374). Cham: Springer International Publishing.

[99] Chesterman, S. (2023). All Rise for the Honourable Robot Judge? Using Artificial Intelligence to Regulate AI.

[100] Shah, N., Bhagat, N., & Shah, M. (2021). Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention. Visual Computing for Industry, Biomedicine, and Art, 4(1), 9.

[101] Belk, R. (2021). Ethical issues in service robotics and artificial intelligence. The Service Industries Journal, 41(13-14), 860-876.

[102] Raaijmakers, S. (2019). Artificial intelligence for law enforcement: challenges and opportunities. IEEE security & privacy, 17(5), 74-77.

[103] Moses, T., & Obi, H. E. (2022). Review on Automobile Crime Prediction Model. International Journal of Darshan Institute on Engineering Research and Emerging Technologies, 11(1), 08-13.

[104] Dakalbab, F., Talib, M. A., Waraga, O. A., Nassif, A. B., Abbas, S., & Nasir, Q. (2022). Artificial intelligence & crime prediction: A systematic literature review. Social Sciences & Humanities Open, 6(1), 100342.

[105] Ekanem, D. (2020). Artificial Intelligence as a Mechanism for Crime Control in Nigeria: A Critical Appraisal. LAP LAMBERT Academic Publishing.

[106] Marr, B (2022). The 5 Biggest Tech Trends In Policing And Law Enforcement. Forbes. Retrieved from https://www.forbes.com/sites/bernardmarr/2022/03/08/the-5-biggest-tech-trends-in-policing-and-law-enforcement/?sh=118c6cff3840

[107] Redden, J., Aagaard, B., Taniguchi, T., & Criminal Justice Testing and Evaluation Consortium. (2020). Artificial Intelligence in Law Enforcement. U.S. Department of Justice, National Institute of Justice, Office of Justice Programs. http://cjtec.org/

[108] Lau, T. Predictive Policy Explained. Retrieved from https://www.brennancenter.org/our-work/research-reports/predictive-policing-explained

[109] Butt, M (2024). Met Police arrest 13 people in two days using ‘dystopian’ facial recognition cameras. Retrieved from https://www.independent.co.uk/news/uk/home-news/met-police-croydon-facial-recognition-b2484112.html

[110] Chen, H (2023). ‘Like something out of Black Mirror’: Police robots go on patrol at Singapore airport. Retrieved from https://www.cnn.com/2023/06/18/asia/police-robots-singapore-security-intl-hnk/index.html

[111] Government of Dubai (2023) Dubai Police Implements AI in Planning and Analysing Traffic Accidents. Retrieved from https://mediaoffice.ae/en/news/2023/October/19-10/Dubai-Police-Implements-AI-in-Planning-and-Analysing-Traffic-Accidents

[112] Mozur, P et al (2022) An Invisible Cage': How China Is Policing the Future. Retrieved from https://www.nytimes.com/2022/06/25/technology/china-surveillance-police.html

[113] Jany, L (2023). LAPD to use AI to analyze body cam videos for officers' language use. https://www.latimes.com/california/story/2023-08-22/lapd-to-use-ai-to-analyze-body-cam-videos-for-officers-language-use

[114] Fares, N. Y., & Jammal, M. (2023). AI-Driven IoT Systems and Corresponding Ethical Issues. In AI and Society (pp. 233-248). Chapman and Hall/

[115] Ballantnye, N (2023). The harm that data do: The case of PredPol. Retrieved from https://medium.com/@neilballantyne/the-harm-that-data-do-the-case-of-predpol-17603c59a1e2

[116] STE Staff (2022). Why a Cybersecure Video Surveillance System is Critical. Retrieved from https://www.securityinfowatch.com/video-surveillance/article/21268413/why-a-cybersecure-video-surveillance-system-is-critical

[117] Bhuiyan, J (2023). First man wrongfully arrested because of facial recognition testifies as California weighs new bills. Retrieved from https://www.theguardian.com/us-news/2023/apr/27/california-police-facial-recognition-software

[118] Perle Systems (2023). Beyond the Badge: Exploring the potential of AI in law enforcement. Retrieved from https://www.perle.com/articles/beyond-the-badge-exploring-the-potential-of-ai-in-law-enforcement-40196867.shtml#:~:text=Facial%20recognition%20and%20identification,AI%20applications%20by%20law%20enforcement

[119] CanLII Canadian Legal Information Institute. Retrieved from https://library.upei.ca/canlii-canadian-legal-information-institute_database#:~:text=Funded%20by%20Canada's%20lawyers%20and,provincial%20and%20territorial%20law%20societies.

[120] Holak, B (2019). NYPD's Patternizr crime analysis tool raises AI bias concerns. Retrieved from https://www.techtarget.com/searchbusinessanalytics/news/252459511/NYPDs-Patternizr-crime-analysis-tool-raises-AI-bias-concerns

[121] Smith, M and Austin, R (2015). Launching the Police Data Initiative. Retrieved from https://obamawhitehouse.archives.gov/blog/2015/05/18/launching-police-data-initiative