IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 18 June 2023, accepted 25 June 2023, date of publication 20 July 2023, date of current version 2 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3294569

SURVEY

A Review of Trustworthy and Explainable Artificial Intelligence (XAI)

VINAY CHAMOLA[1,2], (Senior Member, IEEE), VIKAS HASSIJA[3], A RAZIA SULTHANA[4], DEBSHISHU GHOSH[5], DIVYANSH DHINGRA[5], AND BIPLAB SIKDAR[6], (Senior Member, IEEE)

[1]Department of Electrical and Electronics Engineering, BITS-Pilani, Pilani Campus, Pilani 333031, India
[2]APPCAIR, BITS-Pilani, Pilani Campus, Pilani 333031, India
[3]School of Computer Engineering, National University of Singapore, Singapore 119077
[4]School of Computing and Mathematical Sciences, University of Greenwich, SE10 9LS London, U.K.
[5]Department of Computer Science, Jaypee Institute of Information Technology, Noida 201309, India
[6]Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077

Corresponding author: Vinay Chamola (vinay.chamola@pilani.bits-pilani.ac.in)

This work was supported in part by the Ministry of Education, Singapore, under Grant A-0009040-00-00 and Grant A-0009040-01-00.

ABSTRACT The advancement of Artificial Intelligence (AI) technology has accelerated the development of several systems that are elicited from it. This boom has made the systems vulnerable to security attacks and allows considerable bias in order to handle errors in the system. This puts humans at risk and leaves machines, robots, and data defenseless. Trustworthy AI (TAI) guarantees human value and the environment. In this paper, we present a comprehensive review of the state-of-the-art on how to build a Trustworthy and eXplainable AI, taking into account that AI is a black box with little insight into its underlying structure. The paper also discusses various TAI components, their corresponding bias, and inclinations that make the system unreliable. The study also discusses the necessity for TAI in many verticals, including banking, healthcare, autonomous system, and IoT. We unite the ways of building trust in all fragmented areas of data protection, pricing, expense, reliability, assurance, and decision-making processes utilizing TAI in several diverse industries and to differing degrees. It also emphasizes the importance of transparent and post hoc explanation models in the construction of an eXplainable AI and lists the potential drawbacks and pitfalls of building eXplainable AI. Finally, the policies for developing TAI in the autonomous vehicle construction sectors are thoroughly examined and eclectic ways of building a reliable, interpretable, eXplainable, and Trustworthy AI systems are explained to guarantee safe autonomous vehicle systems.

INDEX TERMS Artificial intelligence (AI), trustworthy AI (TAI), eXplainable AI (XAI), autonomous vehicles, healthcare, IoT.

I. INTRODUCTION

Artificial Intelligence (AI) is a technology that has been growing considerably over the years. It has grown to the extent where it can beat humans in open challenges and has become a need for most humans in their everyday lives [1]. An exemplary case of this would be the strategy game of Go where AI AlphaGo beat the human world champion Lee Sedol in 2016 [2]. This accompanies the fact that AI has applications in almost every field. Be it self driving cars, smart assistants, recommendation engines, disease detection, or automated robots, people's lives are greatly influenced by AI breakthroughs in a variety of sectors [3], [4], [5], [6]. This spurt does not desist here. According to the International Data Corporation, investment in AI is anticipated to increase from $37.5 billion in 2019 to $97.9 billion in 2023. The worldwide AI software industry is prognosticated to flourish in the imminent years, with a market value of over $126 billion by 2025 (Fig. 1, published by Statista Research Department) [4].

Trust is a belief that gives a person the assurance that whatever they put their trust in would not cause them any form of harm. Trust when breached, results in misuse, abuse and

The associate editor coordinating the review of this manuscript and approving it for publication was Zijian Zhang.

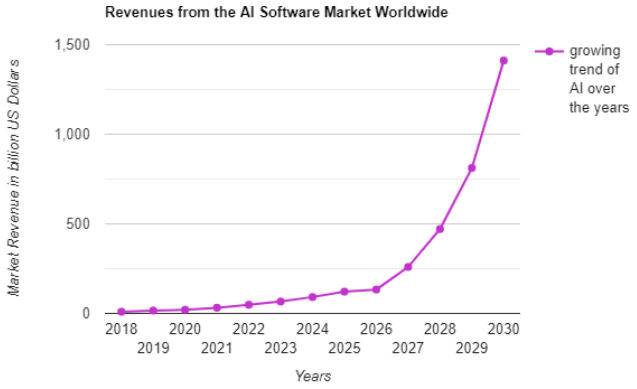

FIGURE 1. Worldwide AI revenue and growth.

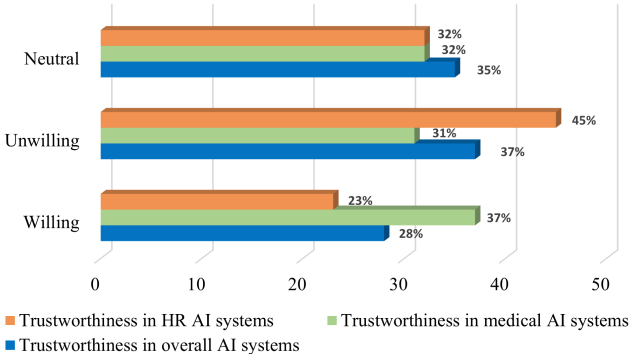

FIGURE 2. Trustworthiness in AI systems.

disuse of the trustee. Primitive machines and algorithms did not have trust issues as the machine would perform whatever task it had been coded to perform, in a way it was the most absolute form of trust. A machine incapable of autonomy was not a threat and could be easily trusted and people could easily believe more in a machine output rather than human output. The advent of Machine Learning (ML) and the concept of making machines think and perform tasks autonomously, however, has resulted in a breach of the trust that existed, as machines can now operate and act independently [7]. ML has enabled devices to operate in a way like never before, wherein it can perform human like tasks in a much faster and accurate pace than any human potentially can. In the current world, image analysis, speech analysis, large scale data analysis and other important tasks are generally done with the help of ML. These tools enable large businesses, medical professionals, and many scientific research fields to make new technologies and to identify flaws in many sectors that were never identified before. Any computational system and even our daily lives are highly facilitated by complex ML algorithms and AI that have changed how we live. The inner workings of most of these algorithms, however, are not understood and are taken for granted. Hence, while they are reliable at the moment, their evolution scale is extremely fast and soon these systems might be capable of deception. Studies have shown that the capabilities of current deep learning algorithms are nowhere near to the maximum potential of any AI. Hence, the amount of trust that is given to any AI as of now is highly unjust and excessive [8].

This transgression, if not realised properly, can result in major issues in the upcoming future. Hence, there are a myriad of concerns associated with the rapid development and dissemination of AI. They vary from the potential of invading people's privacy or following them unwillingly through the Internet via the ClearView AI, to the prevalence of racial prejudice in commonly used AI-based systems, to the quick and uncontrolled generation of economic losses via autonomous trading agents [9]. Future AI might be capable of deceiving the entire human race into doing something that would be catastrophic for them and may get enough resources and potential to conquer humans. To avoid such complications and to ensure that people can freely use the powers of AI without worrying, a certain degree of trust has to be proven by fair means [9]. The form of AI which can be considered trustworthy is called a Trustworthy Artificial Intelligence (TAI). A study by the University of Queensland, Australia in 2021 [10] provides details about the level of trustworthiness of people over overall AI systems, AI in medical and human resource systems. The data are consolidated and shown in Fig. 2. Due to the fact that more than 70% of individuals have opted to have neutral or no confidence in AI systems, the total proportion of trustworthiness in AI systems is just a quarter. It is also found from the study that the public is more trusting and supportive of AI use in healthcare.

A data protection and privacy legislation with 99 articles and 173 recitals was released by the European Union in May 2018 with regard to managing and processing personal data [11]. The articles include legal requirements that an organization should follow and recitals include any other supporting information. Then, in June 2020, a study on General Data Protection Regulation (GPDR) on AI was done, citing GDPR [12], and it states that AI should be compliant with GDPR laws.

The requirements for attaining full trustworthiness are numerous. However, the European Commission (EC) has attempted to justify the criteria in their own means. The EC provides a set benchmark for AI to be evaluated on in order to ensure a level of trustworthiness towards it. The EC has provided seven different criteria on the basis of which trustworthiness can be proven [13], [14] and these rules were formulated on the basis of three main criteria which are to ensure the transparency, reliability and for the protection of data in models [14]. Seven main rights have been highlighted from the impact of GDPR (Table 1) whose violations have stirred the EU to develop the policies for TAI, which have led to the genesis of the criteria. The following is a list of the criteria and their briefing:

- The first of the highlighted criteria is human agency and oversight and it deals with the protection of fundamental rights of humans and proper human AI interaction. This condition has been created to ensure that AI does not

TABLE 1. Seven principles of trustworthy AI.

Seven Requisites for Achieving Trustworthy AI			
All applicable laws and regulations, as well as a set of requirements, should be respected by Trustworthy AI. Specialized evaluation lists attempt to verify the implementation of each of the essential requirements.			
Principles	Explanation	Rights	GDPR Ref
Human Agency and Oversight	Human agency and basic rights should be supported by AI systems, rather than limiting or misguiding human autonomy	Right to obtain human intervention	Recital 71, Art 22
Robustness and Safety	Systems must be safe, trustworthy, and resilient enough to deal with mistakes or inconsistencies	Right to contest a decision in solely automated decision making	Art 22
Privacy and Data Governance	Citizens should have complete control over their personal information, and information about them should not cause any damage or harm to them	Right of notification and access to information about logic involved in automated processing	Art 13, Art 14, Art 15
Transparency	Artificial intelligence systems should be traceable and transparent	Right to obtain an explanation	Recital 71
Diversity and Fairness	AI systems should take into account the whole spectrum of human calibers, skills, and needs, as well as ensuring accessibility	Right not to be subject to solely automated decision making	Art 22
Societal and Environmental Well-being	AI systems should be utilized to promote positive social transformation as well as environmental sustainability and accountability	Right on information of significance of and potential effects of solely automated decision making	Art 13, Art 14 Art 15
Accountability	Mechanisms should be put in place to guarantee that AI systems and their outputs are held accountable	Right to be notified of solely automated decision making	Art 13, Art 14

go against human intentions and works only for human assistance rather than harboring ill intentions.

- Robustness and safety in an AI ensures that the AI is consistent and can rectify its errors and inconsistencies. This handles the category of faulty AI systems.
- Privacy and data governance ensure that the information used by AI does not leak to other sources or harm the user of AI. The personal information of a user should be protected from any malicious activity and be safe from harm.
- The AI should transparent, that is, it should be explainable, traceable and communicable in a way that the creator and the user are able to understand the functioning of the AI.
- Diversity and fairness are an important factor. The AI should be able to understand the differences in humans and should not discriminate against users on the basis of generated stereotypes. It should be an impartial system and abide by the law.
- Societal and environmental well-being have to be maintained by AI. The creation of AI should not cause any harm to the environment and should be a sustainable system reliant on energy sources that are renewable.
- Accountability of AI would ensure that AI can report any issues found in its system and be responsible for them. Mechanisms should be set in the AI so that it can be held accountable for any problems in the data or the output it generates.

These criteria will be elaborated in later sections and their uses and developments in respective fields will be recorded.

There exist several surveys conducted that primarily focus on one or two aspects of TAI. For example, the relationship between trust in AI and trustworthy AI technologies is given in the Joint Research Centre (JRC) regulations by the European Commission [14]. A survey on the robustness of AI-based prognostic and systems health management is presented in [15]. The European Commission Joint Research Centre's technical report intends to aid in the development of a robust AI regulatory framework. It offers an objective assessment of the present state of AI, with a focus on robustness and explainability [14]. A survey of Data-Driven eXplainable AI (XAI) has also been conducted that delineates the major milestones in XAI development as well as highlights a comprehensive taxonomy for expansion and evaluation of XAI [1]. To the best of our knowledge, no comprehensive study has been conducted on all elements of trustworthy AI, as well as the integration of XAI with various sectional backgrounds. This review study aims at addressing the gaps in the literature by giving a comprehensive assessment that spectates all the elements of TAI and XAI from development to evaluation, and spotlights some of the most recent breakthroughs and improvements that have been

achieved towards trustworthy as well as explainable AI. The extensive overview, challenges, and perspectives toward TAI and XAI in different sectors (banking, healthcare, IoT, and, independent AI) are the quintessential emphasis of this survey paper.

The organizational structure of this paper is as follows (Fig. 3): Section II brings light to all the aspects and concepts of TAI and gives a detailed overview of them.

Section III brings forth a brief review listing the need for TAI, achieving Explainability with AI/ML systems. Section IV gives a glimpse of the current status, advancements and developments of TAI and XAI in divergent sectors and fields. Section V sums up the challenges in current streams and highlights a perspective on achieving TAI for various organizations. Further in this survey, future goals and needs, and how trust and explainability is the pressing priority for upcoming AI systems have been covered in Section VI, followed by the concluding remarks in Section VII.

II. AN OVERVIEW OF TAI

ML approaches dominate AI today, with the fundamental feature of building a reasoning system directly from data, typically in massive amounts, without explicit rules to obtain the process outcome. These approaches are particularly generic which make them appealing for a wide range of applications. Furthermore, the ML community has always taken an open approach to cooperation and dissemination, with a wide collection of resources, ranging from software to data sets to documentation, made freely available to everybody. This technique increased the appeal of ML in the scientific and technical communities, as well as its acceptance by practitioners in a variety of fields, by using the massive quantity of data gathered in digital systems (dard report, trust report, security and privacy for AI).

Artificial intelligence has been demonstrated to have a ''black box'' syndrome, owing to absence of insights into how systems operate, prompting implications about obfuscations, arbitrary prejudice, legitimacy, and implications to human confidentiality. This lack of candor is frequently accompanied with underlying biases and inclinations.

TAI is a substantially vast concept with the aim of making AI easier and safer to use than ever before. Hence, the difference aspects of TAI have to deliberated individually and their aspects have to be discussed. This section focuses on the seven different aspects of TAI (Fig. 4) and each topic is addressed in detail.

A. HUMAN AGENCY AND OVERSIGHT

According to the premise of concern for individual autonomy, AI systems should promote human autonomy and decision-making. This necessitates AI systems functioning as facilitators of a democratic, vibrant, and egalitarian society by fostering basic rights and promoting user agency, as well as allowing for human monitoring. Current AI systems have very little access for AI monitoring and challenging its decisions [16]. This is a major issue as AI is ubiquitously used in our daily lives and hence, having no means to challenge its decisions and providing them too much autonomy can result in malpractice by AI. AI being able to perform such tasks unbeknownst to humans is an issue that needs to be monitored and addressed. Hence, human agency and oversight over AI is important to keep the AI systems in check [14], [16], [17], [18].

1) HUMAN AGENCY

Users must be able to make autonomous, enlightened conclusions on AI systems. They should be provided with the information and equipment needed to perceive and communicate with AI systems to a reasonable level, as well as the ability to adequately self-assess or critique the system. AI systems should encourage people to make smarter, more rational decisions based on their performance objectives. AI systems may occasionally be used to shape and influence human behaviour through techniques that seem to be difficult to identify because they use sub-conscious processes such as unethical persuasion, deceit, swarming, and indoctrination, all of which potentially jeopardise personal sovereignty. Hence an active agency should be established with individuals who are highly knowledgeable on the AI and its functioning to monitor it. This agency can be local as well as global. Local agencies would be the ones managed by the creator of AI and hence all large AI creators, such as Google and Meta, who have created an AI used by masses should have a human oversight system in place for their AIs. A global agency also needs to be established to ensure that the AI made by smaller creators are handled properly. The global agency would also be responsible for monitoring the larger creators on certain occasions so that no human malpractice is done by them [14], [16], [17].

2) HUMAN OVERSIGHT

Human oversight guarantees that AI systems do not compromise human sovereignty or have certain detrimental ramifications. Governance tools such as a human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) methodology could be used for monitoring and surveillance. Human intervention throughout every decision phase of the system, which is often neither feasible nor preferable, is alluded to as HITL. The capability for human interference during the system's conceptual stage and overseeing its functioning is referred to as HOTL. HIC refers to the ability to supervise the AI system's entire activities and also the authority to decide when and how to employ it in any particular circumstance [16], [17], [18]. A detailed description of such a system can be found in the article written by Fanni et al. [16].

B. ROBUSTNESS AND SAFETY

Robustness is the quality of being strong and reliable. In the case of an AI, robustness indicates the AI's ability to constanly provide results accurately. For example, in the case of language translations, humans unarguably perform better and

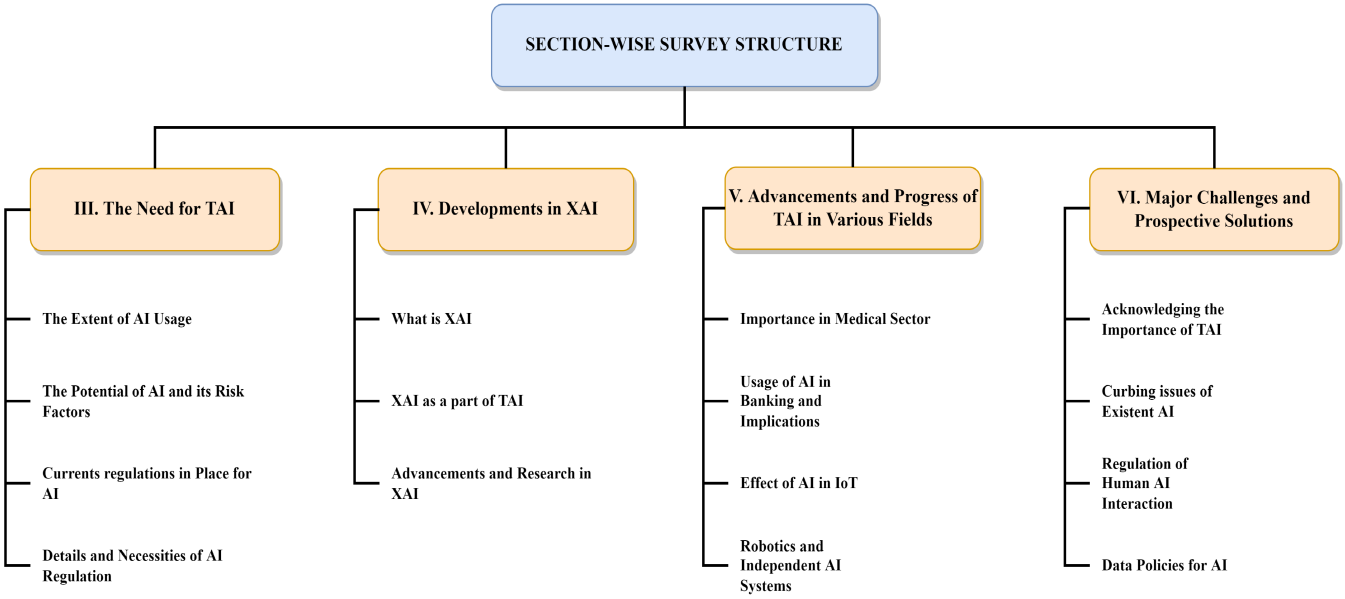

FIGURE 3. Survey organization.

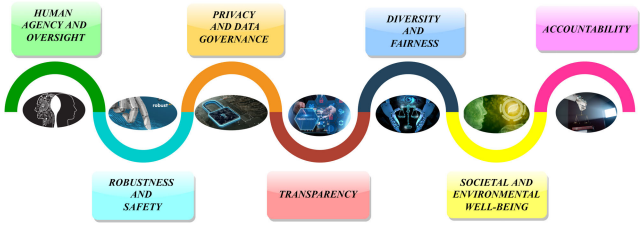

FIGURE 4. Different aspects of Trustworthy AI.

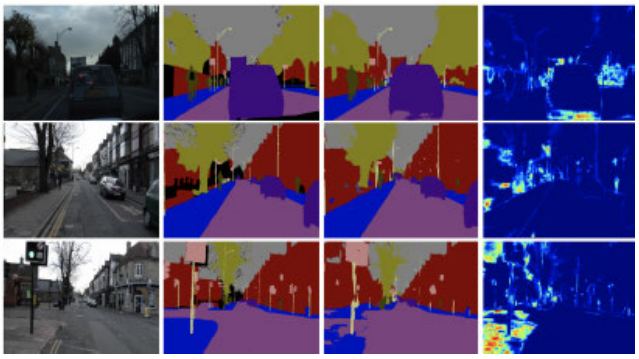

FIGURE 5. Uncertainty in computer vision.

provide a much better explanation of the language compared to most AI translators [19]. Technical robustness, which is inextricably tied to the tenet of risk mitigation, is a fundamental component of ensuring trustworthy AI. Technical robustness necessitates the development of AI systems that take a proactive stance to risks and that dependably operate as intended while mitigating unintended and unforeseen harm and preventing catastrophic harm. This should also apply to prospective changes in their operational environment, as well as the inclusion of other agents (both humans and machines) who might engage with the system in an antagonistic way [15], [19].

Creating ML systems that are resilient to adversarial instances is one of the most pressing current concerns in AI safety. Robust ML systems must be able to acknowledge data that varies substantially from training data and respond against adversarial attacks. A diverse spectrum of research disciplines are striving to make progress in this approach. Incorporating predicted uncertainty estimations into ML systems is one such research path. In this method, each forecast made by the system would be accompanied by a probability estimate. A human operator can be notified whenever the predictive model expresses skepticism about the veracity of its prognosis [15], [19], [18].

A predictive ambiguity estimate for autonomous cars is illustrated as an example (Fig. 5). The image fed into the system can be seen in the first column, the ground truth classification of objects in the image (buildings, sky, street, sidewalk, etc.) can be seen in the second column, the model's classification is seen in the third column, and the system's uncertainty about its classification is seen in the rightmost column. The system is dubious about its identification of sections of the sidewalk, as seen in the figure on the bottom right, and might warn the human operator to take over the guiding wheel [20].

In certain cases, AI has reached a level of robustness that is unmatched by any human talent such as in the case of AlphaGo which beat Lee Sedol, who was the world's top ranked player at the time. In that tournament, Lee beat AlphaGo only once in the five matches [2]. Deepmind later

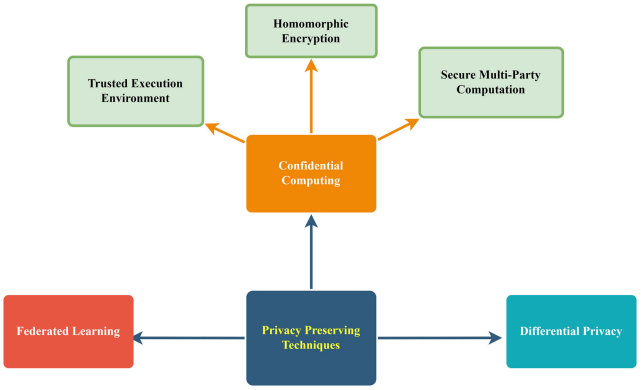

FIGURE 6. Privacy preserving techniques: Overview.

went on to build more such deep learning based AI such as the AlphaGo Zero which learnt games like Atari, Chess, Go and Shogi from scratch and later MuZero, which is the primary version of the AI for all the games. These last two AI systems are fully autonomous with no human intervention and hence, such an AI system being robust is very important. Any instance of rule breaking by such AI in other sectors such as for translation and other human uses could result in damages [2], [19], [18].

C. PRIVACY AND DATA GOVERNANCE

Privacy, a basic right particularly imperiled by AI systems, is inextricably associated with the idea of harm prevention. Mitigation of security loss also entails efficient data stewardship, which encompasses the authenticity and integrity of the data employed, its significance in perspective of the realm wherein the AI systems will be implemented, access procedures, and the ability to interpret data in a way that safeguards privacy. Systems which employ distributed learning are highly susceptible to privacy breach and data leaks as they use data from shared systems to train the algorithm. This allows the organisation to access personal human data through the use of the application. The entire basis of tailor made advertisements for every person is one use of such distributed AI [21]. Since privacy is a much more serious issue in current times than before, regulations are being implemented to employ stricter rules and enhance privacy. GDPR has already stated this issue and countries like Malaysia have also employed some regulations [14], [22]. This is also a major issue in the case of medical data wherein the information in a patient's files can be leaked [3], [23].

To develop trustworthy AI systems, the confidentiality of private and sensitive information conveyed by data and models which may be disseminated across the AI system must be safeguarded. Few privacy preserving techniques are shown in Fig. 6. Considerable emphasis has been placed on the conservation and enforcement of data privacy. For example, the California Consumer Privacy Act (CCPA) was passed into law in 2018 to strengthen privacy rights and regulatory frameworks in California by offering customers more authority over the personally identifiable information that businesses collect and the European Union has enacted the GDPR to ensure data security by granting ownership over the collection and use of private details.

D. TRANSPARENCY

Trust is a person's conviction in the competence and dependability of another person or object. Thus, having a thorough understanding of and familiarity with the system is one of the most crucial elements in building trust [24], [25]. This is where the idea of transparency for a reliable AI first appears. Transparent AI aims to provide for thorough explanation and communication of an AI model's output [14].

As AI systems become increasingly powerful and gain new abilities, AI systems are now increasingly complex. This level of complexity needs deeper explanation to be understood by a normal citizen. Hence, this calls for better explanations and highlighting any aspects that might hamper humans [14], [24]. Any human should have the right to know what the AI's features entail and help with, along with its shortcomings. This would in turn help in ensuring trustworthiness of the AI and fulfilling the right to information for any citizen and protecting the rights of the users.

Having a transparent AI would help to keep the AI in check and free of issues as any underlying issues in the AI would be visible to everyone. A human agency can be used to maintain such records of the AI and assist in proving that the company creating the AI is not falsifying the reports. Multi-level proof systems can be used to maintain such an organisation and this would work as an independent body which would oversee AI operations.

E. DIVERSITY AND FAIRNESS

AI are programs which are created by humans for other humans. Humans are generally inherently biased beings and we tend to discriminate against other things or even humans which do not resemble us. This is hence a trait that can be passed on to AI. It may not be a deliberate act by humans but there are examples of AI making statements against certain humans in a clear depiction of racial behaviour. In the long term, if self conscious AI are not curbed of this behaviour, it could lead to serious issues.

AI which is biased against humans would inherently be a partial being. A new AI with partiality in its coding can start off by being partial towards different types of human beings which is terrible in itself, but with consequent evolution might be discriminating its services against humans as a whole [7], [12]. With the eventual creation of sentient AI, this would lead to even greater problems as AI starts to hide information from humans, their creators. Hence, it is important that AI remain neutral and impartial to all beings to prevent any form of future catastrophes [14].

This is a topic that has not seen much research and hasn't been treated with the amount of importance it should be given as AI are still, mostly, fully supervised under human control.

This, however, needs to change as AI becomes more powerful and with the creation of AI which have achieved human levels of intelligence, this issue needs to be addressed with more concern and given priority.

F. SOCIETAL AND ENVIRONMENTAL WELL-BEING

Throughout the AI system development cycle, global society, other sentient entities, and the ecosystem should all be considered as stakeholders [14]. AI systems' sustainability and environmental accountability should be stimulated, and exploration into AI solutions addressing global concerns, such as the sustainable development goals, should be encouraged. AI systems should ideally be leveraged to benefit all beings, including subsequent generations [12].

AI systems have the potential to contribute to the resolution of some of society's most urgent issues, but this must be done in the most environmentally friendly approach conceivable [26]. In this regard, the system's design, implementation, and use procedure, as well as its extensive distribution network should be scrutinized, for instance, through a detailed appraisal of resource consumption and energy usage during training, with fewer detrimental alternatives considered. Measures to guarantee that AI systems' overall distribution network is environment conscious should be fostered [18].

From changing climate, susceptibility to nuclear proliferation, and zealotry, global challenges are becoming progressively complicated in terms of synchronization. This implies that they can only be remedied efficaciously if all stakeholders co-design and co-own the alternatives and cooperate together to institute them. With its data-intensive, algorithmic-driven solutions, AI may significantly support the grappling of such logistical intricacy, leading to greater communal coherence and cooperation.

G. ACCOUNTABILITY

The notion of fairness is inextricably tied to the imperative of accountability. It entails implementing organizational processes to assure accountability and responsibility for AI systems and their consequences, both before and after they are developed, implemented, and used [27], [28]. Accountability is a relational condition; it can't be described as an agent's characteristic in isolation from other agents. The following three aspects are present in each and every description of accountability.

1) RESPONSIBILITY

The responsibility for one's deeds and decisions, which also acts as a framework for moral acclaim or censure, communal approbation, and the possibility of legal ramifications.

2) ANSWERABILITY

There are two dimensions to it: (i) the competence and Willingness to communicate with a designated counterpart for the reasons behind actions and (ii) entitlement of such counterpart to ask for such grounds to be divulged.

3) SANCTIONABILITY

Sanctionability represents the ability to act against the accountable and monitoring party. Issues between the aforementioned needs may emerge during implementation, culminating in inescapable trade-offs. Within the latest advancements, such trade-offs should be handled in a pragmatic and structured manner. This necessitates that the AI system's pertinent objectives and ideals be specified, and if a contradiction emerges, trade-offs be explicitly acknowledged and assessed in terms of their danger to ethical standards, including fundamental human rights. The design, implementation, and usage of the AI system must not progress in that form if no morally acceptable trade-offs can be established.

III. NEED FOR TAI

As mentioned in the introduction, AI usage is constantly on the rise. The revenue spent on AI and it being constantly implemented on mobile devices and everyday use items, has made AI an important segment in our daily lives.

A. AI IN HEALTHCARE

AI is one of the most important developments in medical fields where it is estimated that in a few years, AI has the potential to address 20% of all clinical needs [29]. Medicine is a field of study which influences the most number of people. In the field of medicine, AI is being used in almost every part, from drug manufacturing to surgical procedures.

In the case of drug development and discovery, several complex models are used to simulate different compounds and their potential reactions to the human body. They are also used in monitoring clinical trials, and data gathering and analysis from them [30]. AI is also used for prognosis and diagnosis of patients in big hospitals. A major application of AI in medicine is image analysis in radiology, and AI being increasingly adept at gathering information from images and finding underlying patterns that even humans can't identify has become a very powerful tool in this field [31]. Surgery is also a field where AI is being researched and trained to perform remote surgeries with the help of robotics. AI can now perform invasive surgeries with extreme precision that humans could never do before with their hands [32]. In all these sections, AI has proven to be quite a game changer with the introduction of new mechanics and additions which allow robots and humans to perform tasks that could not be achieved before. This also further enhances our reliance on robotics, AI and machinery. This reliance is also a case of giving more power into AI's hands [25].

Medicine is the key to the current longevity of humans and is arguably the key to life and death for the modern world. In such a scenario, AI needs to be highly precise and robust. Any errors by AI or the human involved in the creation of the AI, could lead to loss of lives [30], [33], [34]. In the event where AI can infiltrate and alter databases of hospitals, or produce erroneous results of diagnosis, the hospitals, their patients and the families of the patients will be adversely

affected. In such an event, the usage of AI and the reliance on it will cause irretrievable losses and hence this is an issue that should be addressed immediately.

B. AI IN BANKING AND COMMERCE

Banking is another important sector where AI and AI based systems are being used. These systems handle massive amounts of data and keep all our data safe and secure from prying eyes. It can be easily seen and attested that no banking system in the current age and date is possible without the involvement of data science. Data science is the study of data, and facilitates data-enabled or data-guided products which might include discoveries, predictions, services, recommendations, insights into decision-making, ideas, models, paradigms, tools, and systems [35].

Data science is a field where the scientist extracts useful information from a given piece of data. Organisations in the current world use this discipline to analyse, predict, and accurately work in the direction which enhances their productivity [35], [36]. The AI used in banking can be referred to as AutoAI, which are essentially automated systems used to clean, label and enhance the data overall. These models ease the work of a data scientist and at times can even assist in helping to create more useful statistics which were previously unseen by the naked eye. These AutoAI models are sandbox type systems which can be modified and suited to individual requirements [36], [37], [38]. Hence, these are in use by wide variety of people across the globe. There exist more advanced versions of AutoAI systems which can be used to predict and create insights which hold the power to change the paradigm of the entire company, hence making it superior to others. One such system is AutoAI-Time Series Forecasting (AutoAI-TS), which uses a complex set of pipelines and time series statistics to predict complex results. This system was made by IBM, and used by Sunganthi et al. and many others to predict different medical diseases which are nearly impossible to predict by humans alone [36], [37], [39], [40]. In the world of banking, however, this translates to predicting the market fluctuations using the statistical data of organisational and world trades.

This would hence lead to the question ''What happens when an AI powerful enough to perfectly predict the entire market comes by?'', and another question would be, ''Who has the possession of that AI?''. These two questions themselves raise concerns regarding the power that a single AI system could hold. Anyone possessing such powerful AI would basically have the ability to calculate the best possible profit at any given time. An organisation possessing this would hence have the power to stump their entire competition, becoming a monopoly by just controlling their and other's assets [25]. With the current technology in existence, such an AI is a nearly impossible feat as it would require a extremely large supercomputer which would have to hold almost all the trade details of the entire world, many Zettabytes of information, and have the computational power to process all data in an instant. However, this cannot be said about the future. Technology is constantly progressing and one day some organisations might be able to achieve such a feat. Hence, to prevent such an AI from being abused, proper rules should be made for any AI system that may be created.

C. AI IN IoT

Among all the important places that AI exists, the one place where AI usage is the highest is the Internet. The Internet of Things (IoT) is a network system of wired, wireless or sensor based based things. The Internet is system which upon its introduction, has been accepted by the masses at incredible speed. This major development in the modern world has not only helped the daily life of a common man but is also a system used by all major industries in managing, creating and constantly monitoring their activities. The Internet is also host to various algorithms and various artificially intelligent mechanisms that help humans and organisations in daily life. It has also become a place which is host to a huge amount of data. This data, however, is also prone to attacks and in the hands of a malicious individual can result in global issues.

AI mechanisms use massive amounts of data to first train themselves in order to give meaningful results for a targeted issue. Hence, they are bound to be in connection with that data. If an AI system is somehow hacked into, it can become a funnel of leaking data which can be used very easily by adversaries. Data leaks from major companies and organisations have become a common event right now. Traditional AI uses data from a centralised data collection system. These data centers gather all the data and provide necessary information to the AI. The vulnerability for this AI is that the pools of data centers, if hacked into, can lead to massive data leaks. One development that avoids these data data centers is the concept of federated learning. Federated AI has been created for a distributed environment where every device can host an AI [41]. While this is a good step towards privacy in TAI, this AI has become a much more complicated system. Privacy is still a major concern in this AI as it uses huge amounts of data and depending on how the data is used and security systems in place, the data is still at risk.

A system being hacked into is one thing, however, having an AI which can leak information on its own or use the information it gets for malicious purposes is a different thing. This is the reason why the robustness of AI is very important in every scenario. An AI being robust would safeguard people from the information being leaked. The next most important factor, for an AI on the Internet, is to ensure that it provides equal results for every human. AI cannot be allowed to discriminate and learn human discrimination principles. Having an AI that can discriminate against humans is a recipe for disaster that, in the long term, can potentially lead to AI turning against humans. If an AI is not equal for all then a sentient AI will most definitely discriminate against humans. Once a sentient AI adopts humans principles and realises that they are superior to humans, it could lead to all AI turning against humans.

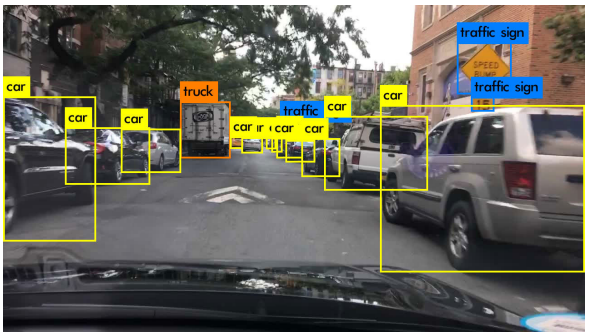

FIGURE 7. Predicting the object's behaviour: explainable and interpretable.

D. ROBOTICS AND AUTONOMOUS AI

The usage of AI in robotics is quite obvious. They are the key components that are enabling two legged robots and systems which are trying to mimic humans, to exist. For such creations, the first duty of AI would be to help humans in physical tasks. They are made to assist humans and work for humans. The AI responsible for such creations should be very robust, so no one can interfere with its functioning and work without a specific set of controls.

E. EXPLAINABILITY IN AI

The term 'explainability' refers to the process of making the technology/system understandable to human and how easily they comprehend the model's results. Also, it is when they are aware of whether the outcomes are accurate or require a second judgment. One way to achieve XAI is to build interpretable learning algorithms [42]. There is a thin line that separates an interpretable ML algorithm from an explainable one. Although an explainable algorithm enables you to understand every node's functionality, an interpretable algorithm aids in understanding the rationale behind every output. Being interpretable will also make an algorithm explainable. For instance, with self-driving cars, if we can track each car's behaviour (i.e., if each car's behaviour can be clearly explained in many scenarios), we can very accurately predict the reason behind generating the output (Fig. 7). Alternately, more accurate models are difficult to interpret and building an explainable model is also equivalent to developing an interpretable one.

The adoption of interpretable ML models, such as Regression and Decision Trees (DT), can eliminate the need for black box models; although, doing so requires comes at a cost. For instance, a combination of input features with varying weights determines the output feature in linear regression. Nevertheless, employing only regression is unlikely to produce effective results when features are highly correlated. DT, on the other hand, tries to build hierarchical relationships between the features and classifies them based on a threshold value. Regression demands that the variables have a direct or inverse connection, whereas DT do not. Yet, the location of the feature in the tree's elevation decides how much the feature is weighted and any change in the features characteristics causes a big change in the prediction result. DT are the replica of if-then rules and a number of works like RuleFit [43] which employs DT have proved capable of understanding the sparse relation between the variables which is quite difficult to understand by regression models. Another category of interpretable models is known as Generalized Additive Models (GAMs), that employ a flexible function known as a spline to understand the relationship between the non-linear data. This smoothing function, termed as a spline, leverages the coefficients of a linear regression model to create not only linear curves but also wiggly curves for evaluating and interpreting data points. The GAM is formed by the combination of splines.

Another way to achieve XAI is to build a model using agnostic methods [44]. These models isolate the explanations from the ML black boxes. The majority of ML algorithms that can be explainable are anticipated to be agnostic models. An agnostic model is adaptable to employ any ML method and offers flexibility in both its explanation and representation. It does, however, make it simple to compare algorithms. There are two methods for putting the agnostic models into practice: employing global analysis and local analysis. The local technique analyzes on each prediction results, whereas the global method analyzes the models overall output. The global methods include Partial Dependence Plot (PDP), Accumulated Local Effect Plot (ALE), H-Statistic, Global Surrogate models and work on combination of features/overall understanding of the data, whereas the local methods include Individual conditional expectation curves (ICE), Local Surrogate Models (LIME) etc, and they discuss how an individual feature contributes to the prediction or can be combined to form a global surrogate model.

The PDP shows the effect of two or more features on the prediction variable. It very clearly demonstrates whether or not there is a linear relationship between the input and output variables. When it plots the highly correlated features, the PDP plots exhibit significant bias. Nonetheless, the ALE plots handle strongly correlated characteristics better than PDP since they are less susceptible to them. On the contrary, ALE are complex and difficult to implement than PDP. The H-statistic (feature interaction) demonstrates that the prediction does not depend on the sum of all the features in the feature set because each feature's contribution varies and there is inter-feature dependency. For example, consider working with a dataset that includes information on the location, size, and market value of a house. Since the size and property value depends on the location of the house, one approach to predict the property value is to measure the interaction between the features. Also, a two-way relationship between the features may be examined by establishing a partial dependency function as they interact among themselves.

The ICE plots are for local prediction analysis, which contrasts PDP. PDP does not plot every instance but instead shows the average influence of the features on the output whereas ICE plots every instance of change against the

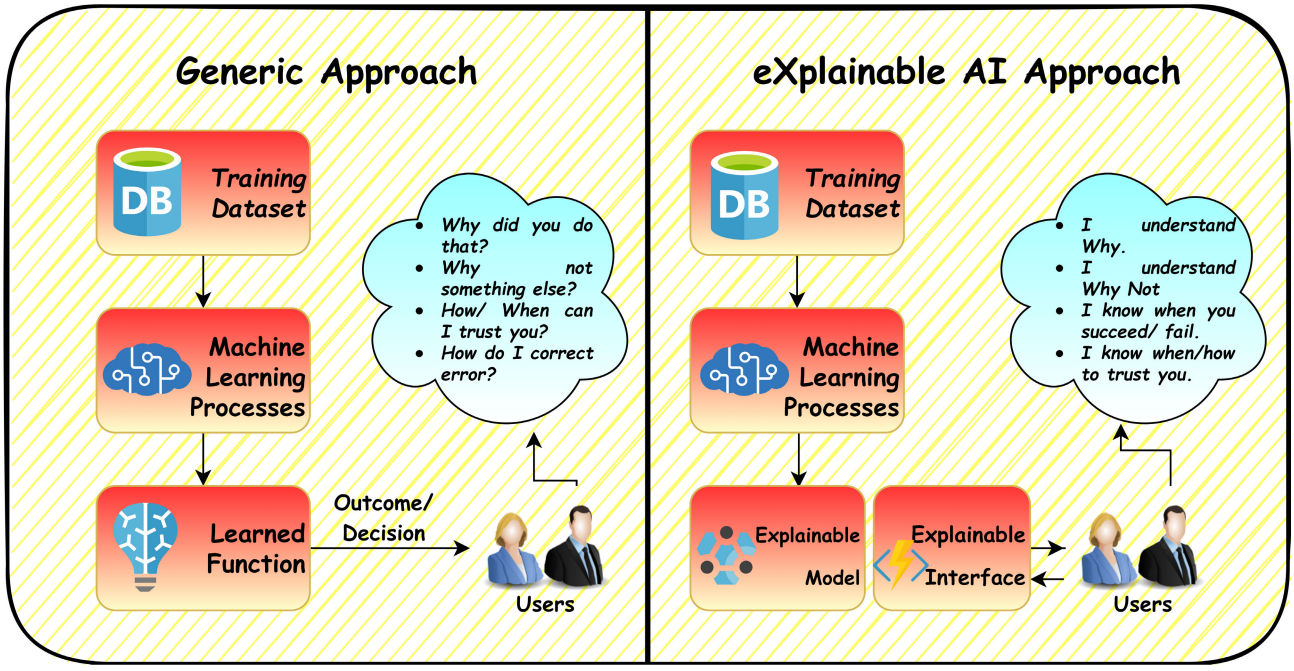

FIGURE 8. AI today and tomorrow with XAI.

output. LIME demonstrates how individual predictions in black box models affect the overall output. By altering the input data points, LIME compares the output of the original and modified data points and determines how the model behaves. LIME also measures the closeness of the outcomes of the original data points and the modified ones. The use of model-agnostic methods to create interpretable models is growing in popularity in recent years. These methods, apart from automatically tuning the hyperparameters and creating ensemble and stacked models, also understand the importance of the features, and plot partial relationships. The requirement is not only to analyse the data but the models. An interpretable ML model creates an explainable one and, as a result, develops into a trustworthy one. A self-explanatory robot response is crucial, not just for the questions it answers but also for the algorithms it employs and the procedures it takes to evaluate the data. An AI that can be understood is one that has programmes that can self-explain themselves.

IV. DEVELOPMENTS IN XAI

as AI develops, humans are under pressure to understand and replicate how an AI algorithm arrives at a result. The entire calculation process is transformed into a ''black box,'' which is incredibly challenging to understand [28]. There is a need for even more transparent systems that can explain their results due to the prevalence of black-box models in most private and industrial AI/ML systems that employ deep learning and other machine learning approaches. In the words of the US Defense Advanced Research Projects Agency, XAI is ''AI technology that can express its logic to a human user, identify its strengths and limits, and impart knowledge of how it could perform in the future'' [28]. It depends on explainability to create a system that is reliable and understandable. Explainability must be successfully achieved to increase public trust in the computational execution. The AI in use today and the one expected to work with tomorrow is shown in Fig. 8. If a system should be held accountable, steps should be taken to ensure that input inefficiencies are addressed and reduced [45], [46].

The models are black boxes due to their significant non-linearity and intricacy, making it unfeasible for individuals to fathom their fundamental operational processes and judgment mechanisms [28]. Such obscurity might cause severe problems and stymie future AI advances. To commence with, it is difficult to challenge a black-box model's decision-making. Passengers, for example, may feel tremendously apprehensive about the self-driving system if the projections are not self-explanatory, such as when the car abruptly turns left at the crossroads when it ordinarily continues straight without explanation [27], [45]. Furthermore, black-box models are challenging to govern and preclude anomalous behavior. DNNs have been considered ineffective and potentially exploited by hostile perturbation, a major concern in AI reliability. Leading to significant biases and prejudices in the training set, black-box models may still make poor predictions even in the absence of adversarial attacks [28], [46], [47]. The implications in essential situations, like a medical diagnostic, might be calamitous or even lethal. The explainability of AI models, or the capacity to acquire insights into the mechanisms underpinning model dynamics, is widely desired to address these challenges. Heading back to the self-driving scenario, if the car turns left and reports, ''There are a car collision 500 meters in front of us'' then the passengers will believe and accept the autonomous assessment that ''turning left will take us

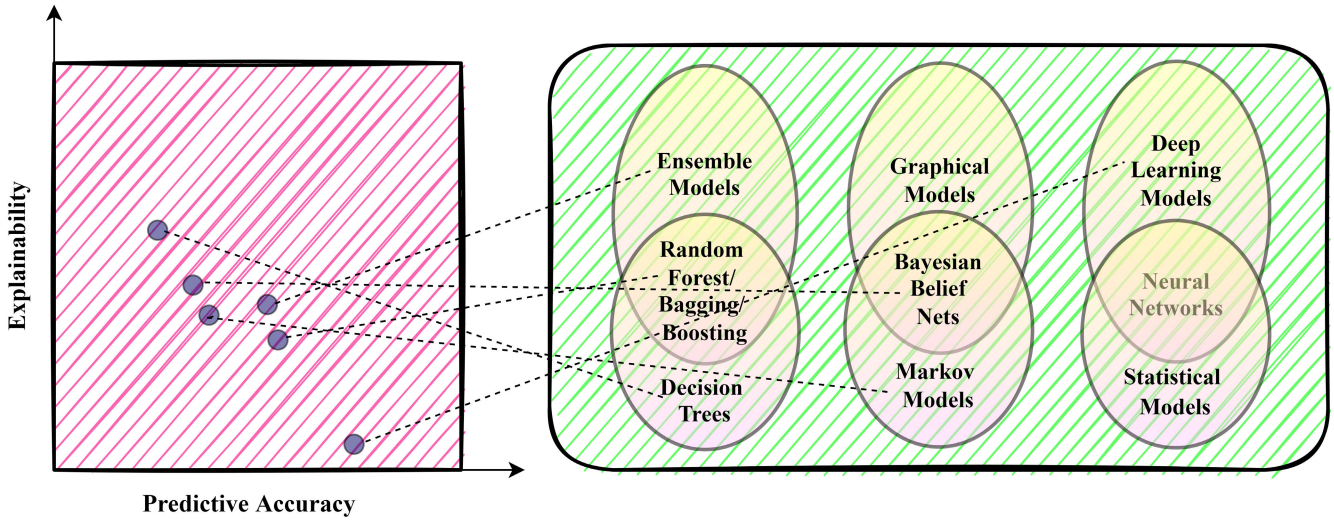

FIGURE 9. Machine learning models' explainability tends to be inversely correlated with their predictive performance.

10 minutes longer than anticipated, but heading straight will take probably 40 extra minutes'' [4].

The reliability and predictive accuracy of an AI model is often inversely correlated with its interpretability as shown in Fig. 9; the better the predictive accuracy, the less explainable the model [45]. The DARPA XAI research offers a chart to exemplify this intriguing paradox, demonstrating that while SVMs, ensemble models, and decision trees have the greatest explainability levels among the indicated ML approaches, they have the poorest predictive performance [12]. On the other hand, deep neural networks are least likely to be comprehensible while having the best predicting ability of any learning technique.

In recent times, AI researches are focused on deciphering the enigma of neural network models and develop a transparent architecture. Transparency design and post-hoc explanation are the two key areas of exploration in XAI (Fig. 11). The transparency structure divulges how a model functions [48], [49]. It makes an effort to comprehend the conceptual framework (Fig. 10), including how a decision tree is built, discrete components, such as a logistic regression criterion, and training procedures, such as stochastic optimization that solutions seek. According to users, the post-hoc explanation illustrates why a result is extrapolated. It endeavors to provide analytical assertions, including why a commodity is suggested on a shopping website, provide visualisations, such as a transmission map that underlines the prominence of each pixel in the categorization and segmentation of an entity [4], [49].

A. TRANSPARENT DESIGN

Transparent systems are intended to be comprehensible and interoperable. The ultimate level of transparency is retained by simulable models, next by decomposable structures, and subsequently by computationally transparent configurations [46], [50].

Simulatability is the capability for user simulation: ''A paradigm where a person can take in incoming data along with the model's attributes and proceed through each computation required to construct a prognosis in an acceptable timeframe'' [28], [50].

Decomposability refers to the capability to dissect a model into its component elements (inputs, variables, and operations), and then to explicate every one of these components. This is the second degree of transparency. Since all of the input variables have to be easy to comprehend, articulating all of the system's components and operations is a barrier when trying to break the system down into its constituent aspects [45], [51].

Computational Transparency represents the third level transparency and the competence to comprehend the steps the model goes into producing its outcome. The said property is satisfied, for instance, by models that categorise instances based on certain measure of similarity (such as K-nearest neighbours, K-means, and SVM), as the methodology is intuitive: locate the datum that would be the most parallel to the one presumed, and allocate the aforementioned to the very same category as the latter [48]. On either side, complex loss functions are generated by complicated models, such artificial neural networks, and the training goal solution also has to be anticipated. Generally, the sole prerequisite for a model to be included in this category is that the user must be able to evaluate it using a statistical and parametric analysis [52].

B. POST-HOC EXPLAINABILITY

Post-hoc explainability methodologies establish lesser intricate surrogate models to emulate complicated black-box ML techniques [1], [50]. To comprehend and interpret the internal dynamics of black-box ML techniques and, consequently, the real-world solicitation of models' post-maneuver, human auditors can leverage these simplified surrogate models.

By interrogating the network and developing a white-box proxy model, post-hoc explainability accepts a trained classifier as an input and uncovers the dynamic linkages that the framework has acquired. Post-hoc explanations transform facts from a massive, comprehensive structure (the black-box model) into a simplified, compact one through a procedure termed as model condensation (the white-box surrogate model) [53].

Formerly, post-hoc explainability yielded two distinct ''classifications'' of interpretability: global and local. Global interpretability illustrates a model's overarching rationale and the justification for all probable consequences [54]. Global model interpretability is the explanation through the architecture and characteristics of a network and illustrates a system using the most salient rules uncovered from the training set [1], [53]. For a specific prognosis, local interpretability addresses model properties as well as the relevance of input information. Local models expressed as a sequential function of input variables can be more efficient than model simulations because limited segments of the system are more inclined to be continuous [52], [55].

1) GLOBAL METHODS

Global techniques seek to offer comprehensions of a model's rationale and the full justification for all the forecasts, based on the comprehensive perspective of its peculiarities, learning aspects and hierarchies, etc. There are a myriad ways to investigate global interpretability [1], [54]. We segregate them into the following three subcategories for better readability: ***Model extraction*** is the process of partitioning an interpretable prototype from the original black-box model. ***Feature-based methods*** are used to guesstimate the symbolic importance or appropriateness of a component, and ***transparent modeling approach*** is the process of altering or revamping a black-box model in order to make it more easily decipherable [28].

a: MODEL EXTRACTION

Model extraction's fundamental tenet is to train an explainable system, sometimes alluded to as a global surrogate model, that imitates the characteristics of a black-box system. One may use the procedures described further to create such surrogate prototype: Choose dataframe D that can be a novice one or one of the training sets of black box frameworks [51]. Then, apply a black box analysis to the specified sample D and acquire the related forecasts. Decide on a kind of decipherable paradigm and build the required model. Thus, employ dataset D to develop the explainable model to meet black-box forecasting accuracy. Amongst the most prominent designs for model extraction is indeed the decision tree since it is typically regarded as basic and comprehensible. However, a decision tree becomes difficult for humans to comprehend when there are too many nodes or it is too huge. Since no particular black box prototype is expected all throughout the training process, the structure is model-agnostic [28], [48],

FIGURE 10. A conceptual framework of the explanation process in the context of XAI.

[1], [56]. Hinton et al. present knowledge distillation as a unifying approach for model extraction instead of constructing methods for extracting a specific model. By optimizing the loss function underneath, the fundamental premise is to develop a simplified prototype system to emulate the complicated primary model [1], [57]:

$$Loss = \sum_{i=1}^{n} p_i \log y_i(T = 1) - \sum_{i=1}^{n} x_i(T = t) \log y_i(T = t)$$

where p_i denotes class i's real world label, and [1]

$$y_{i(T)} = \frac{e^{\frac{m_i}{T}}}{\sum_j e^{\frac{m_i}{T}}}.$$

b: FEATURE BASED METHODS

The community has embraced the elegant and concise option of extending universal insights through an extracted paradigm. Nevertheless, such extraction puts emphasis on compressing the dimensionality of the problem, which undermines the veracity of the initial formulation [1], [57]. In the burgeoning domain of data mining, feature selection is a preprocessing strategy for efficient and scalable analysis that attempts to identify a subset of original characteristics in order to lessen the feature map as quickly and efficiently as possible even while accomplishing the predefined criteria. The genetic algorithm has been shown to be a remarkably optimal technique in a multitude of issues necessitating near-optimum searches. To determine a subset of parameters that are most significant to the categorization job, a novel hybrid algorithm is used in [50]. There are two phases to the optimization technique. Together, the medial and lateral

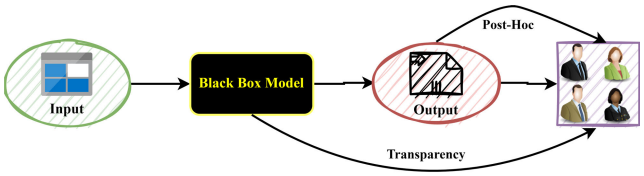

FIGURE 11. Two main types of eXplainable AI work: post-hoc explanation and transparency design.

optimizations yield greater local search effectiveness as well as growing global model accuracy.

Liu et al. devised an optimized feature extraction technique by incorporating MSPSO, SVM, and the F-score strategy to further resolve the feature extraction challenges [58]. A modified version of particle swarm optimization, Intelligent Dynamic Swarm (IDS) was suggested as a metaheuristic optimization mechanism. Naive Bayes, SVM, and ELM classifiers are applied to 10 datasets from the University of California at Irvine (UCI) ML repository in order to gauge the classification performance of IT-IN as well as the other multi-scale feature selection procedures. To assess the significance of an input characteristic, there are two different work pathways which include feature importance and activation maximization [46], [53]. The influence of each characteristic, perhaps locally or globally, towards the prediction model is evaluated by feature importance. The permutations feature relevance, which was initially developed for random forests, appears to be the most exquisite method. A model-agnostic variant of the notion is subsequently developed by Fisher et al., and is given by [1], [58]:

$$PFI_d = \mathbb{E}(P(\widetilde{f}(\widetilde{X_d}, X_s), Y)) - \mathbb{E}(P(\widetilde{f}(\widetilde{X}), Y)).$$

2) LOCAL METHODS

Local methods make a concerted effort to validate the model's actions for a specific circumstance or cluster of occurrences [54], [59]. Intricate models for local behaviors may be coupled to minimal characteristics by an unambiguous association (such as linearity), which lessens the complexity of interpreting black-box models. Within certain geographical regions, simple functions can help to facilitate factual justifications [48]. We differentiate local approaches into two kinds based on the methods used to derive rationalisations: ***local approximation*** and ***propagation-based methods***.

V. ADVANCEMENTS AND PROGRESS OF TAI IN VARIOUS FIELDS

The ongoing digitization of various industries has embraced AI systems as a crucial component. However, in tightly controlled domains like IoT systems, avionics, finance, banking, autonomous systems, and medical services, integrating AI and its interoperability with legacy applications is much more onerous [27], [47]. Organizations in these sectors need to be cognizant of the constantly shifting regulatory frameworks that would make or break AI ventures since data privacy and security are fundamental, alongside human welfare and safety.

AI is being used in a multitude of sectors. However, in this survey, we emphasize on the applicability and integration of trustworthy AI in the fields of healthcare, banking/finance, autonomous human centric systems, and IoT systems where breakthroughs in AI infrastructure, software, services, and platforms are opening up a plethora of opportunities [60]. In these domains, data validation, reliability, privacy standards, norms for decision-making mechanisms, and comprehensive legislation are constantly changing. Service companies and solution architects may now play a significant role in this ecosystem, with AI adoption yielding a substantial return on the investment [59]. Enhancing efficiency and productivity, establishing effectiveness of the system, information and view analysis, and sentient inference are some of the standout features of AI throughout this domain. The automation of protocols and workloads, enhanced conformity, interfaces with greater consistency and reliability, autonomous pastiche technologies (in finance and banking), and contemporary digital assistants are additional possibilities [53], [61].

Data protection, expense, safety, reliability, assurance, and enhanced decision-making mechanisms are the driving strengths underpinning AI. From the viewpoint of commercial uptake and procedures, each component has a distinctive influence and importance [18]. The operator makes sure that services have a commercial purpose, are compliant with standards, and are intimately associated with both ecosystems. Furthermore, the controllers facilitate widespread adoption sans jeopardizing the terminal interface, which will boost efficiency and productivity [62], [63].

A. ACCOMPLISHING TRUST IN HEALTHCARE SYSTEMS

Amongst the most new emerging markets for the application and implementation of AI systems is medicine and healthcare. With the aid of AI, ailments may be diagnosed quickly and genetic abnormalities that might plague us in the future can be identified. Similar to workplace processing, administrative support computation may be expedited by algorithms, enhancing patient interaction whilst economising by minimizing inefficiencies. Such savings are then used to enhance the quality of care. To guarantee that development is for the betterment of everybody, technology in the medical industry should indeed adhere to the law, standards, and federal regulations [29], [64].

We categorize explainable and interpretable AI paradigms according to the necessity they aim to achieve since this acknowledges the overall objectives of modeling and analysis criteria, and consequently affects the design decision of comprehensible AI technologies. We consider the following three aspects of explainability and interpretability:

1) TO FACILITATE THE VALIDATION AND ASSESSMENT OF OTHER MODEL REQUIREMENTS

Healthcare operations are extensively multifaceted, making it challenging to conflate and quantify all essential characteristics (including legitimacy, integrity, and resilience) in

a framework by utilizing conventional preliminary testing grading rubrics. Explanations may assist in allowing individuals to suggest remedial measures in specific instances. For example, healthcare professionals seek to examine whether the analysis incorporated (in)appropriate attributes. Whilst the use of interpretability to validate alternative model demands and expectations is quite often acknowledged with in literature and research, we emphasize that justifications never ensure that system categorical imperatives are achieved [65], [66].

2) TO MODERATE INTERPERSONAL COMMUNICATION

The social aspect of justifications can serve as an incentive for the necessity of explainability. Establishing a comprehensive understanding of the decision-making procedure is among the main triggers behind why individuals frequently seek for clarifications and explanations. Compliance with the ''privilege to explanations'' underneath the European Union GDPR is indispensable in the healthcare sector. Professionals ought to be willing to defend their actions to their counterparts and clients, even if there is no statutory necessity to do otherwise [12], [14].

3) TO GAIN FRESH PERSPECTIVE

In order to better understand from the frameworks for information retrieval, one might likewise incorporate predictability. Explainability improves learning for educational and recreational purposes via permitting contrasts of learnt approaches with previously acquired learning and knowledge. Extensive research can indeed be impacted by these breakthroughs, for instance, by using them to develop and inject drugs or organize medical testing [1], [34].

On the premise of the process that generates the explanations, the explanation's nature, its purview, the paradigm type it may illustrate, or an amalgamation of these criteria, numerous ontologies have already been postulated. We emphasize on model-independent methodologies for post-hoc explanations. Starting with authenticity, model-based interpretations invariably meet the thoroughness criterion since they present enough features to determine the consequence for a specific intake. Pertaining to explainable accounting, coherence is met whenever the work description is employed as the rationale. We may evaluate the reliability of post-hoc model-based solutions using the bitrate indicator proposed by Markus et al. [4] and Lakkaraju et al. [67]. They quantify the fraction of comparable forecasts and determine authenticity as the extent of (dis)agreement between the scheduling process and post-hoc model justification. In accordance with understandability, we contend that although global explanations provide a reasonable rationale for the entire paradigm, they often meet the transparency feature. As a corollary, strategies that rely on global models, attribution, and examples are clear.

Recently, as ML has garnered prominence in the healthcare industry, several complaints and concerns have also surfaced. Proposed approaches that use U-Net for clinical categorization are some of the most effective approaches. U-Net, nevertheless, is a black-box system and not easily decipherable since it is a deep learning neural network. Furthermore, several domain-especial techniques and unique modifications have been developed [68], [69], [70].

U-Net is an extensively adopted Fully Convolutional Network that has been employed to mediccal and healthcare visual segmentation and consists of an encoder, a bottleneck component, and a decoder. Due to its U-shaped topology coupled with semantics, rapid development pace, and minimal bandwidth consumption, U-Net addresses the needs of diagnostic visual categorization. Pharmaceutical and diagnostic scans frequently incorporate various segments and are thus pixelated inside a volumetric dimension. To appraise a 3-dimensional picture, a 2-dimensional analysis technique is frequently employed. A 3-dimensional U-Net prototype developed from a 2-dimensional U-Net is intended to focus upon structures with various shapes and sizes. The system trained using Attention Gate inferentially trains to suppress superfluous areas in such an input patch and accentuate eye-catching elements appropriate for specialized activities. This makes it much harder that conspicuous external vasculature segmentation components of cascading Convolution Neural Network will be used in foreseeable time. The model's responsiveness and accuracy are strengthened by integrating Attention Gates with a Convolutional Neural Network structure like U-Net. Zhou et al. presented U-Net++, a new, broader neural net topology for image processing, to take categorization further and farther [68], [70], [71].

Numerous quatrefoil techniques can be employed to develop reliable AI in the healthcare industry:

4) EXTENSIVE DATA PRESENTATION

Actual statistics may indeed be skewed, erroneous, or inadequate since they is not always gleaned for scientific objectives. Since it permits comprehensive insights into the constraints of the AI framework and characterizing the system performance, data presentation can be as significant as interpretability and explainability. Markus et al. [4] outline an extensively recognized methodology that could be employed to examine if Electronic Health Records (EHR) data is adequate for a certain application instance. This paradigm assesses the quality of data focused on concordance, coherence, and credibility and may be employed to present the conclusions to clients of the AI system designed with the content in an organized manner.

5) THOROUGH VALIDATION IS CARRIED OUT

By employing evaluative feedback, questions related to the reliability or comprehensiveness of systems may be resolved. Due to the absence of accuracy and consistency, recreating the forecasting models on new data can be a time-consuming procedure. The Observational Health Data Sciences and Informatics consortium has designed tools that facilitate the establishment and performance evaluation of diagnosis

estimation techniques at mass, in a transparent manner, and in compliance with acknowledged guiding principles [72], [73]. Potential options to produce highly reliable AI include exploring methods for including model demands and expectations throughout model optimization, and establishing quality control for such system criteria.

6) ORDINANCE AND REGULATIONS

Regardless of the fact that governance of AI systems remains in the preliminary stages of research, proven oversight over other protection applications, including medication management, implies that it might eventually prove to be a viable method for fostering credibility. An approach to monitoring the final outcome would be to introduce uniform development criteria that need to be adhered to in order to regulate the design process. To oversee researchers, we may establish a regulatory system, as proposed by Nazar et al. [74]. This system permits for fiduciary ethics and therefore is analogous to specialists within medical industry holding credentials. The US Food and Drug Administration is presently contemplating new guidelines for digitization, one of which involves a switch inside the locus of control from finalized outcome to enterprises [65].

B. ACCOMPLISHING TRUST IN BANKING AND COMMERCE SECTOR

Among the main impediments prohibiting banking institutions from implementing their AI strategy is the ''black-box'' paradox [28]. The burgeoning domain of XAI may give banks more insight and specificity on existing AI regulations while assisting them in navigating challenges of trust and openness. The objective of XAI is to augment the explanation, usability, as well as comprehension of AI frameworks without jeopardizing the overall effectiveness or precision while making predictions. Furthermore, explainability is a growing issue for financial institutions who seek to ensure that AI conclusions and procedures are ''easily comprehensible'' by bank officials. Various client advocacy organizations, competitors, and internal stakeholders within banking firms all exhibit this upsurge [1], [51].

There are several additional tangible perks that an efficient XAI strategy may provide for enterprises and firms. Depending on the queries being addressed as well as the modelling methodologies being employed, explainability techniques may disclose a variety of details regarding a particular system. For instance, XAI methodologies that demonstrate how well an approach works might well be effective for deciphering the connections between variables, determining why a system is performing inadequately, or uncovering possible privacy violations. Together, such endeavors seem to be crucial for maintaining consumer rights as well as equitable loaning because they assist in determining hyperparameters that seem to have contrasting effects, comprehend exchange in system performance, create stronger proposals for model adaptation. also, they facilitate organizations to gain better confidence and reliability with their configurations, and guard against probable legal or regulatory hurdles [75], [76].

The World Economic Forum claims that ML and AI will revolutionize the financial industry by delivering novel approaches for users to interact with financial organisations [75]. Instances of existing AI usage scenarios throughout the finance market proffered by the Dun and Bradstreet report encompass: ''enhanced virtual assistants, authenticity confirmation in consumer rollout, invoice data analysis, anomaly detection throughout reimbursement, cost structure in investing, anti-money embezzlement surveillance, price adjustment in vehicle insurance, automated assessment of official documentation, customer service, regulatory compliance, asset management, trading implementation, and equity operational processes''. The Financial Regulation Authority and the Bank of England assert that the incorporation of learning algorithms in UK financial institutions is surging, and it is anticipated to more than double in the next three years. ML has been most widely employed across client-facing operations (including advertising and client support) in addition to anti-money laundering and anomaly detection. Several businesses also leverage ML in trading cost and implementation, credit risk analysis, and traditional insurance costing and screening [25], [77].

C. ACCOMPLISHING DIFFERENT ASPECTS OF TRUST AND RELIABILITY IN AUTONOMOUS SYSTEMS

System operation, expedited information dissemination, processing of huge volumes of data, working in potentially hazardous environments, operating with much more resilience and tenacity than people, and perhaps even astronomical exploration are all the capabilities and potentials of autonomous systems [45], [78]. Current automated technologies are the culmination of many years of research and development, which have led to advancements in computer recognition systems, responsive systems, intuitive interface design, and sensing automation. We merely have to consider the vehicles we presently commute in to realize the pervasiveness of autonomous technologies in our daily lives. The marketplace for automotive intelligent hardware, technology, and operations will expand from $1.25 billion in 2017 to $28.5 billion by 2025 as per [79]. Current cars leverage AI for a range of features, including intelligent cruise control, adaptable automatic driving and parking, and blind-spot detection, as shown in Fig. 12.

According to Intel's analysis on the anticipated merits of automated cars, the employment of such innovations on roadways would save commuters' annual travel time by 250 million hours and contribute to saving more than 500,000 lives, solely in The United States of America, between the years 2035-2045 [79], [80]. Along with the possibility of enhancing current lifestyle, there exists a substantial public apprehension regarding the trustworthiness such AI technologies. Such concern, that constitutes a considerable downside, is mostly triggered by accounts of current

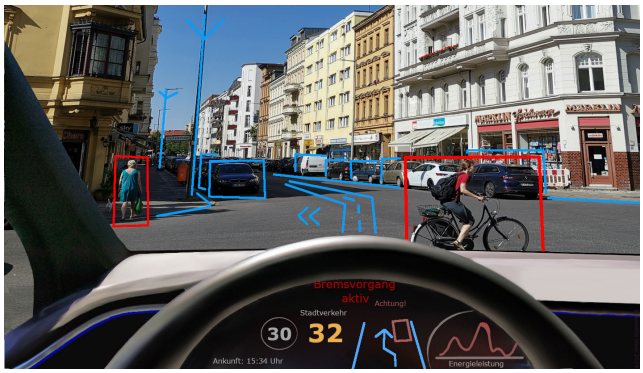

FIGURE 12. An automobile presenting a legitimate and comprehensible justification for its in-the-moment choice, serving as the paradigmatic example of XAI in automated driving.

road crashes using autonomous vehicles, especially because of its improper unilateral judgments.

1) CHALLENGES FACED BY AUTONOMOUS SYSTEMS

Humans are inclined to constantly remain euphoric about the prospects of novel approaches and overlook or apparently seem oblivious of probable pitfalls of cutting-edge advancements. Mankind preferred to endure unreliable commodities and offerings even during early stages of robotics and autonomous systems implementation, but they've progressively recognized that reliable and trustworthy autonomous systems are critical [81]. Countless instances have underlined how trustworthiness significantly impacts the way operators employ automation. Demonstrations of attacks have been carried out to illustrate how automated vehicles might be commandeered, and the malfunction of the Maneuvering Characteristics Augmentation System (MCAS) on Boeing air-crafts led to the crash of two aircrafts which resulted in the fatalities of 157 and 189 passengers, respectively [82], [83]. In circumstances during which they were anticipated to ensure a high level of protection, several autonomous robot systems have tragically faltered. Unexpected and unfavorable occurrences could, in fact, have considerable deleterious effect on how acceptable autonomous robots are. This isn't merely a technological issue since advancements in automated systems also have brought about a multitude of crucial and complicated ethical issues and presented myriad moral, sociological, and regulatory concerns [84].

2) AUTONOMOUS VEHICLES' NORMS AND GUIDELINES

It is imperative to carefully assess how well automation is administered, considering the complications and rising concerns raised by AI technologies. As a corollary, both locally and globally, governmental organizations have begun to construct regulatory regimes to supervise the overall operation of data driven platforms [85]. Such rules are primarily aimed at ensuring that stakeholders retain access to their information and safeguarding their privacy. For instance, the European Union's GDPR framework, which was been passed in 2016 and went into effect in May 2018, established criteria to support the ''entitlement of an explanation'' premise for users [47].

Subsequently, a slew of groups have developed rules to regulate automated vehicles and ensure that they comply with enforcement agencies. Nine criteria have been proposed in the National Association of City Transportation Officials' automated vehicle manifesto as a framework for upcoming autonomous driving governance (Fig. 13) [50], [79]. The R&D Corporation's guidelines are yet another series of policies that address the prospects and challenges of driverless cars, in addition to the linkages between such technology and legal and accountability concerns. Their basic tenets also give comprehensive instructions to regulatory agencies regarding how to conduct detailed investigation into mass transit mishaps and provide safety advice to regulatory bodies and autonomous vehicle production companies like the National Highway Traffic Safety Administration (NHTSA), Society of Automotive Engineers (SAE) International, and Tesla [79], [86].

Regulations have also been endorsed by the International Organization for Standardization (ISO) which address the pertinent autonomous vehicle challenges. Instances of such regulations include the ISO 21448 criterion, that outlines spatial awareness prerequisites to sustain safety and reliability under the ''Protection of the Intentional Features and functions'', as well as the ISO 26262 criterion, that is referred to as ''Road transport - Operational protection'' and addresses the security of electronic and electrical systems in automobiles. The main standard facilitating the systematic use of autonomous vehicle technology in this context is ISO/TC 204, which delivers an extensive manual on the entire system and infrastructure elements of ITS [45], [79], [80].

3) INTERPRETABILITY AND EXPLAINABILITY IN AUTONOMOUS SYSTEMS

The necessity for explanations and interpretations in autonomous systems is prompted by current challenges, predefined policies and guidelines, as well as cross-disciplinary outlooks and social mores. The subjective necessity for XAI in autonomous systems is still mostly spurred by traffic crashes and safety hazards. However, from a social and technological aspect, the basic notion seems to be that human-sentient design, implementation, and distribution of automated systems is indispensable [87], [88]. The development standards of autonomous vehicles should address the demands of the users and be mindful of their preexisting assumptions and beliefs since humans are the key societal agents and consumers of the technology. Taking into consideration these viewpoints, explainable autonomous systems can aid in the following:

- ***Trust and Reliability:*** Humans instinctively desire assurance and confirmation that transport networks are reliable since reckless and negligent driving may significantly affect the security of both passengers

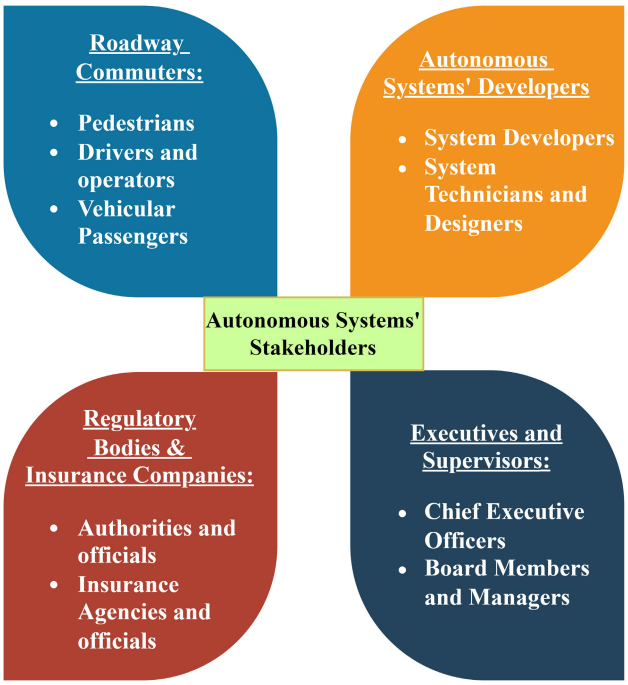

FIGURE 13. Classification of autonomous systems stakeholders.

and pedestrians. Israelson and Ahmed have demonstrated that there exists a fundamental requirement for computational guarantee to foster interactions between human-autonomous systems. Developing explicable autonomous navigation technologies is thus a reasonable requirement for the reliable usage of autonomous systems. The development of reliability will strengthen the engineering behind autonomous vehicles' openness and accountability [27], [89].

- ***Users' Perspective Layout:*** The populace's adoption of the aforementioned technology would expand if the intended consumers' involvement, ideas, and expectations are sought out during the conception and creation of autonomous vehicles [27], [89]. For in-car passengers or standby operators, a self-driving automobile might offer an user interface (UI) to convey the key decisions. There exists a plethora of research studies that employ sentient XAI architectures to relay a vehicle's judgments to the riders and operators within the car via lights, audiovisual, acoustic, and textual information. An essential prerequisite for the general acceptance of autonomous vehicle technology is users' perspective layout that makes use of comprehensible AI techniques. An effective user interaction is established when consumers' and stakeholders' cross-model evaluation is incorporated into the modelled layout of autonomous systems, as demonstrated in a recent empirical research by Atakishiyev et al. [79].

Omeiza et al. have suggested various explanations and interpretabilities in relation to autonomous systems [90]. These can be summarizes as follows.

a: ACCORDING TO TYPE OF CONTENT

In this area, explanations are segmented according to the parts or aspects that system incorporate plus the way they are conveyed. Instances of content categories include interpretations of which system parameters contribute significantly to the predicted responses, input relevance, input susceptibility, testimonials, and socioeconomic considerations.

b: CAUSE-BASED INTERPRETATIONS

Relying upon evidence at hand, explanations leverage predetermined reasons to account for a specific output. Such justifications are produced using cause-based filters, incorporating "W's" questions, for instance, "how did the automobile pick right way rather than the left path?". In reality, it's worth emphasizing that such a form of rationalization is applicable to a wide range of autonomous system scenarios [79], [91].

c: EXPLANATIONS PREMISED AROUND THE SYSTEM CONFIGURATION

This categorization intends to encapsulate the characteristics of the integrated environment by distinguishing between two types of interpretations: information-driven speculations that outline the result of a forecasting model and objective-driven interpretations that outline an agent's actions relying on accomplishing its objective in a pre-configured environment [1], [79].

d: TANGIBLE INTERPRETATIONS

This category encapsulates the viability and variety of justifications that the organization may deliver by just being either local or global. Local interpretations can only represent a certain portion of any and all potential actions, i.e., they can only describe one forecast in a particular circumstance [92]. On the contrary, global justifications are able to justify every major choice made along the way, from the preliminary step to the ultimate stop, including the rationale an autonomous system picked a certain itinerary or altered the intended course mid-trip.

D. EXPLAINABLE AI FOR AUTONOMOUS SYSTEMS

There are many initiatives to create automated systems that produce comprehensible justifications on a automobile's pivotal decisions, spurred by contemporary modeling and analysis of automated vehicle technologies [81]. Textual and visual justifications are ordinarily used to explain the dynamics of automated-driving cars. Knowing the way Convolution Neural Networks record concurrent picture snippets which result in specific vehicular conduct is an essential premise in achieving visual interpretations since DeepNets, usually in reinforced variants like CNN architectures, enable the visual acuity of autonomous vehicles.

In this context, tweaks to produce visual explanations have been implemented as a consequence of comprehensible convolutional networks. Such research also provide impetus for leveraging graphic approaches to explicate autonomous

TABLE 2. Research and surveys on XAI based autonomous systems.

Survey Research Studies	Year Published	Objective	Algorithm(s) Used	Explanation Type	Target Demographic
[96]	2016	Explanations and interpretations of CNN projections contingent on Pixels	Convolutional Neural Networks (CNN)	Visual explanation	Autonomous vehicles' creators
[97]	2019	Visual image comprehension employing semantic segmentation	ENet, SegNet	Visual explanation	Autonomous vehicles' creators
[98]	2019	A comprehensively decipherable neural movement controller	Fast Adaptive Filter (FAF), IntentNet	Visual explanation	Autonomous vehicles' creators
[99]	2020	Explanation of entity response judgments for autonomous systems	Faster R-CNN	Visual explanation	Every stakeholder
[100]	2020	Observation-to-action guidelines that should be internalised for self-driving cars	Long Short Term Memory (LSTM), Mask R-CNN	Visual Textual explanation	Every stakeholder
[101]	2021	Creating causal inferences both in and out of tree-based interpretations	Tree based algorithms	Textual explanation	Autonomous vehicles' creators, regulators
[102]	2021	Goal-based forecasting and scheduling for autonomous vehicles that is comprehensible	Monte Carlo tree search	Textual explanation	Pedestrians, vehicle drivers
[103]	2021	Decipherable purpose identification for autonomous vehicles with obscured parameters	Interpretation of the objective and occluded components, Monte Carlo tree search	Visual explanation	Autonomous vehicles' creators
[104]	2021	Explainable and provable objective identification for autonomous vehicles using trained decision trees	Objective identification using decision trees and explainable trees	Visual Textual explanation	Autonomous vehicles' creators

driving judgements. A methodology for examining how a sequence of pixel intensities influences the CNN's forecasting, called Visual-Back Propagation is presented in [94]. The fore and aft autonomous driving objective tests using Udacity autonomous vehicle database demonstrate the efficacy of the suggested approach for troubleshooting Neural's predictions [99].

A semantic segmentation paradigm that is executed like a pixel-by-pixel categorization has been suggested by Hofmarcher et al. to explicate the fundamental legitimate sense of the surroundings [84]. They utilize CityScapes, a standard data collection for comprehending roads scenes, to assess the viability of their approach. With almost more than half a percent per-class average intersection over union (IoU) and more than 80 per-category average IoU, the methodology surpasses several well-known segmentation models including ENet and SegNet [71], [101]. The model's explainability is beneficial towards unforeseen occurrences since it permits network debugging and clarifies the thinking behind autonomous vehicle choices.

In accordance with the proposal put forth by Zeng et al., autonomous vehicles can be trained to operate securely by adhering to traffic regulations such as relinquishing, interacting with the other pedestrians, and obeying road signs [95]. Researchers employ unprocessed Light Detection and Ranging (LIDAR) data as well as a high definition mapping to provide decipherable outputs including 3-dimensional item recognition, projected future trajectories, and expenditure mapping projections. The system can grasp the operating conditions based on the comprehensive data provided by 3-dimensional recognition occurrences [102]. The L-1 and L-2 ranges used in movement predictions help to determine if faulty movements are indeed the result of inaccurate pace or directional estimations. Furthermore, expense mapping visualization provides a top-down description of the traffic situation [48], [99].

Based in the existing research summarized in Table 2, it can be concluded that XAI for autonomous systems is a catalogue of AI-driven methodologies that (1) guarantee a plausible degree of safety for a vehicle's major decisions, (2) offer rationales and accountability on the intervention choices in crucial traffic situations, and (3) follow all traffic regulations set forth by the law enforcement agencies [79], [103].

VI. CONCLUSION

The objective of this paper is to educate practitioners and scholars on the development and advancement of trustworthy and explainable AI systems in various contexts and to assist in the standardisation of the trustworthy AI discipline. Through the integration of many viewpoints as well as the provision of specific trust metrics and notions, this review offers a comprehensive assessment of the research. Significant constraints still persist despite the tremendous advancements accomplished in explainable AI and intelligence systems. These consist of the replicability of the post-hoc explainability methodologies, the absence of a unified understanding, set of criteria, and set of metrics for the interpretability of intelligence systems, the friction between efficiency and predictability, and the limitations in explaining deep neural networks. The paper provides a summary of the advancements in formalized explainable and trustworthy AI, emphasizes its triumphs, but also discusses its current drawbacks and identifies potential relevant studies. The survey concludes with a description of various methods for accomplishing trustworthy as well as explainable AI systems, and listing the open challenges.

REFERENCES

[1] X.-H. Li, C. C. Cao, Y. Shi, W. Bai, H. Gao, L. Qiu, C. Wang, Y. Gao, S. Zhang, X. Xue, and L. Chen, "A survey of data-driven and knowledge-aware explainable AI," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 1, pp. 29–49, Jan. 2022, doi: 10.1109/TKDE.2020.2983930.

[2] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.

[3] E. T. Joa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021.

[4] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies," *J. Biomed. Informat.*, vol. 113, Jan. 2021, Art. no. 103655.

[5] P. C. M. Arachchige, P. Bertok, I. Khalil, D. Liu, S. Camtepe, and M. Atiquzzaman, "A trustworthy privacy preserving framework for machine learning in industrial IoT systems," *IEEE Trans. Ind. Informat.*, vol. 16, no. 9, pp. 6092–6102, Sep. 2020.

[6] V.-L. Nguyen, P.-C. Lin, B.-C. Cheng, R.-H. Hwang, and Y.-D. Lin, "Security and privacy for 6G: A survey on prospective technologies and challenges," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 4, pp. 2384–2428, 4th Quart., 2021.

[7] Y. Lu, "Artificial intelligence: A survey on evolution, models, applications and future trends," *J. Manage. Anal.*, vol. 6, no. 1, pp. 1–29, Jan. 2019, doi: 10.1080/23270012.2019.1570365.

[8] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *Social Netw. Comput. Sci.*, vol. 2, no. 3, p. 160, Mar. 2021, doi: 10.1007/s42979-021-00592-x.

[9] S. Thiebes, S. Lins, and A. Sunyaev, "Trustworthy artificial intelligence," *Electron. Markets*, vol. 31, no. 2, pp. 447–464, Jun. 2021, doi: 10.1007/s12525-020-00441-4.

[10] N. Gillespie, S. Lockey, and C. Curtis, "Trust in artificial intelligence: A five country study," Univ. Queensland, KPMG, 2021.

[11] EU. *GDPR Articles and Recitals*. Accessed: May 25, 2023. [Online]. Available: https://www.dpocentre.com/resources/gdpr/

[12] (Feb. 2019). *What is GDPR, the EU's New Data Protection Law?* [Online]. Available: https://gdpr.eu/what-is-gdpr/

[13] L. Floridi, "Establishing the rules for building trustworthy AI," *Nature Mach. Intell.*, vol. 1, no. 6, pp. 261–262, May 2019, doi: 10.1038/s42256-019-0055-y.

[14] R. Hamon et al., "Robustness and explainability of artificial intelligence," Publications Office Eur. Union, 2020, vol. 207.

[15] G. Marcus, "The next decade in AI: Four steps towards robust artificial intelligence," 2020, *arXiv:2002.06177*.

[16] R. Fanni, V. E. Steinkogler, G. Zampedri, and J. Pierson, "Enhancing human agency through redress in artificial intelligence systems," *AI Soc.*, vol. 38, no. 2, pp. 537–547, Jun. 2022, doi: 10.1007/s00146-022-01454-7.

[17] J. Anderson, L. Rainie, and A. Luchsinger, "Artificial intelligence and the future of humans," *Pew Res. Center*, vol. 10, no. 12, pp. 1–10, Dec. 2018.

[18] B. Shneiderman, "Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems," *ACM Trans. Interact. Intell. Syst.*, vol. 10, no. 4, pp. 1–31, Oct. 2020, doi: 10.1145/3419764.

[19] T. G. Dietterich, "Steps toward robust artificial intelligence," *AI Mag.*, vol. 38, no. 3, pp. 3–24, Oct. 2017.

[20] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" 2017, *arXiv:1703.04977*.

[21] C. Ma, J. Li, K. Wei, B. Liu, M. Ding, L. Yuan, Z. Han, and H. V. Poor, "Trusted AI in multi-agent systems: An overview of privacy and security for distributed learning," 2022, *arXiv:2202.09027*.

[22] S. Kamaruddin, W. R. W. Rosli, N. N. M. Saufi, A. M. Mohammad, and Z. Hamin, "The quandary in data protection and rights to privacy of AI technology adoption in Malaysia," in *Proc. Innov. Power Adv. Comput. Technol. (i-PACT)*, Nov. 2021, pp. 1–5.

[23] G. Kaissis, A. Ziller, J. Passerat-Palmbach, T. Ryffel, D. Usynin, A. Trask, I. Lima, J. Mancuso, F. Jungmann, M.-M. Steinborn, A. Saleh, M. Makowski, D. Rueckert, and R. Braren, "End-to-end privacy preserving deep learning on multi-institutional medical imaging," *Nature Mach. Intell.*, vol. 3, no. 6, pp. 473–484, May 2021, doi: 10.1038/s42256-021-00337-8.

[24] M. Loi, A. Ferrario, and E. Viganò, "Transparency as design publicity: Explaining and justifying inscrutable algorithms," *Ethics Inf. Technol.*, vol. 23, no. 3, pp. 253–263, Sep. 2021.

[25] S. Sachan, F. Almaghrabi, J.-B. Yang, and D.-L. Xu, "Evidential reasoning for preprocessing uncertain categorical data for trustworthy decisions: An application on healthcare and finance," *Expert Syst. Appl.*, vol. 185, Dec. 2021, Art. no. 115597, doi: 10.1016/j.eswa.2021.115597.

[26] V. Aggarwal and A. Doifode, "Sustainability of CEO and employee compensation divide: Evidence from USA," in *Intelligent Communication Technologies and Virtual Mobile Networks: Proceedings of ICICV 2022*. Springer, 2022, pp. 453–460.

[27] J. Vice and M. M. Khan, "Toward accountable and explainable artificial intelligence part two: The framework implementation," *IEEE Access*, vol. 10, pp. 36091–36105, 2022.

[28] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, Sep. 2019, doi: 10.1145/3236009.

[29] C. González-Gonzalo, E. F. Thee, C. C. W. Klaver, A. Y. Lee, R. O. Schlingemann, A. Tufail, F. Verbraak, and C. I. Sánchez, "Trustworthy AI: Closing the gap between development and integration of AI systems in ophthalmic practice," *Prog. Retinal Eye Res.*, vol. 90, Sep. 2022, Art. no. 101034.

[30] D. Paul, G. Sanap, S. Shenoy, D. Kalyane, K. Kalia, and R. K. Tekade, "Artificial intelligence in drug discovery and development," *Drug Discovery Today*, vol. 26, no. 1, pp. 80–93, Jan. 2021.

[31] A. S. Panayides, A. Amini, N. D. Filipovic, A. Sharma, S. A. Tsaftaris, A. Young, D. Foran, N. Do, S. Golemati, T. Kurc, K. Huang, K. S. Nikita, B. P. Veasey, M. Zervakis, J. H. Saltz, and C. S. Pattichis, "AI in medical imaging informatics: Current challenges and future directions," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 7, pp. 1837–1857, Jul. 2020.

[32] X.-Y. Zhou, Y. Guo, M. Shen, and G.-Z. Yang, "Application of artificial intelligence in surgery," *Frontiers Med.*, vol. 14, no. 4, pp. 417–430, Aug. 2020, doi: 10.1007/s11684-020-0770-0.

[33] T. Grote, "Trustworthy medical AI systems need to know when they don't know," *J. Med. Ethics*, vol. 47, no. 5, pp. 337–338, 2021. [Online]. Available: https://jme.bmj.com/content/47/5/337

[34] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nature Med.*, vol. 25, no. 1, pp. 44–56, Jan. 2019, doi: 10.1038/s41591-018-0300-7.

[35] I. H. Sarker, "Data science and analytics: An overview from data-driven smart computing, decision-making and applications perspective," *Social Netw. Comput. Sci.*, vol. 2, no. 5, p. 377, Jul. 2021, doi: 10.1007/s42979-021-00765-8.

[36] D. Wang, J. D. Weisz, M. Müller, P. Ram, W. Geyer, C. Dugan, Y. Tausczik, H. Samulowitz, and A. Gray, "Human-AI collaboration in data science: Exploring data scientists' perceptions of automated AI," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. 2, pp. 1–24, Nov. 2019, doi: 10.1145/3359313.

[37] S. Y. Shah, D. Patel, L. Vu, X.-H. Dang, B. Chen, P. Kirchner, H. Samulowitz, D. Wood, G. Bramble, W. M. Gifford, G. Ganapavarapu, R. Vaculin, and P. Zerfos, "AutoAI-TS: AutoAI for time series forecasting," in *Proc. 2021 Int. Conf. Manage. Data*. New York, NY, USA: ACM, 2021, pp. 2584–2596, doi: 10.1145/3448016.3457557.

[38] H. H. Hoos, "Automated artificial intelligence (AutoAI)," 2018.

[39] N. Madian, "An approach for predicting heart failure rate using IBM auto AI service," in *Proc. Int. Conf. Comput. Intell. Knowl. Economy (ICCIKE)*, Mar. 2021, pp. 203–207.

[40] J. N. V. R. S. Kumar, K. H. Kumar, A. Haleem, B. Sivaranjani, B. N. T. Kiran, and S. Prameela, "IBM auto AI bot: Diabetes mellitus prediction using machine learning algorithms," in *Proc. Int. Conf. Appl. Artif. Intell. Comput. (ICAAIC)*, May 2022, pp. 24–29.

[41] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. V. Poor, "Federated learning for Internet of Things: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1622–1658, 3rd Quart., 2021.

[42] U. Kamath and J. Liu, *Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning*. Cham, Switzerland: Springer, 2021.

[43] J. H. Friedman and B. E. Popescu, "Predictive learning via rule ensembles," *Ann. Appl. Statist.*, vol. 2, no. 3, pp. 916–954, Sep. 2008.

[44] S. R. Islam, W. Eberle, S. K. Ghafoor, and M. Ahmed, "Explainable artificial intelligence approaches: A survey," 2021, *arXiv:2101.09429*.

[45] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, "Explainable AI: A brief survey on history, research areas, approaches and challenges," in *Proc. CCF Int. Conf. Natural Lang. Process. Chin. Comput.*, 2019, pp. 563–574.

[46] F. K. Dosilovic, M. Brcic, and N. Hlupic, "Explainable artificial intelligence: A survey," in *Proc. 41st Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2018, pp. 1–6, doi: 10.23919/mipro.2018.8400040.

[47] Z. Yang, A. Zhang, and A. Sudjianto, "Enhancing explainability of neural networks through architecture constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2610–2621, Jun. 2021, doi: 10.1109/TNNLS.2020.3007259.

[48] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[49] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[50] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2805–2824, Sep. 2019, doi: 10.1109/TNNLS.2018.2886017.

[51] I. Stepin, J. M. Alonso, A. Catala, and M. Pereira-Fariña, "A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence," *IEEE Access*, vol. 9, pp. 11974–12001, 2021.

[52] M. Nassar, K. Salah, M. H. U. Rehman, and D. Svetinovic, "Blockchain for explainable and trustworthy artificial intelligence," *WIREs Data Mining Knowl. Discovery*, vol. 10, no. 1, pp. 1–12, Oct. 2019, doi: 10.1002/widm.1340.

[53] F. Yang, M. Du, and X. Hu, "Evaluating explanation without ground truth in interpretable machine learning," 2019, *arXiv:1907.06831*.

[54] D. Vale, A. El-Sharif, and M. Ali, "Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law," *AI Ethics*, vol. 2022, pp. 1–20, Mar. 2022, doi: 10.1007/s43681-022-00142-y.

[55] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *J. Artif. Intell. Res.*, vol. 70, pp. 245–317, Jan. 2021, doi: 10.1613/jair.1.12228.

[56] Q. Zhang, Y. Nian Wu, and S.-C. Zhu, "Interpretable convolutional neural networks," 2017, *arXiv:1710.00935*.

[57] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," 2018, *arXiv:1808.00033*.

[58] Y. Liu, G. Wang, H. Chen, H. Dong, X. Zhu, and S. Wang, "An improved particle swarm optimization for feature selection," *J. Bionic Eng.*, vol. 8, no. 2, pp. 191–200, Jun. 2011, doi: 10.1016/s1672-6529(11)60020-6.

[59] N. Frosst and G. Hinton, "Distilling a neural network into a soft decision tree," 2017, *arXiv:1711.09784*.

[60] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a 'right to explanation,'" *AI Mag.*, vol. 38, no. 3, pp. 50–57, Oct. 2017, doi: 10.1609/aimag.v38i3.2741.

[61] S. Elbaum and J. C. Munson, "Software black box: An alternative mechanism for failure analysis," in *Proc. 11th Int. Symp. Softw. Rel. Eng.*, 2000, pp. 365–376.

[62] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, Jun. 2018, doi: 10.1145/3236386.3241340.

[63] X. Renard, N. Woloszko, J. Aigrain, and M. Detyniecki, "Concept tree: High-level representation of variables for more interpretable surrogate decision trees," 2019, *arXiv:1906.01297*.

[64] L. T. Majnarić, F. Babič, S. O'Sullivan, and A. Holzinger, "AI and big data in healthcare: Towards a more comprehensive research framework for multimorbidity," *J. Clin. Med.*, vol. 10, no. 4, p. 766, Feb. 2021. [Online]. Available: https://www.mdpi.com/2077-0383/10/4/766

[65] P. Savadjiev, J. Chong, A. Dohan, M. Vakalopoulou, C. Reinhold, N. Paragios, and B. Gallix, "Demystification of AI-driven medical image interpretation: Past, present and future," *Eur. Radiol.*, vol. 29, no. 3, pp. 1616–1624, Mar. 2019, doi: 10.1007/s00330-018-5674-x.

[66] A. Morales, D. D. Kinnamon, E. Jordan, J. Platt, M. Vatta, M. O. Dorschner, and C. A. Starke, "Variant interpretation for dilated cardiomyopathy: Refinement of the American college of medical genetics and genomics/ClinGen guidelines for the DCM precision medicine study," *Circulat., Genomic Precis. Med.*, vol. 13, no. 2, Apr. 2020, Art. no. e002480.

[67] A. Lakkaraju, A. Umapathy, L. X. Tan, L. Daniele, N. J. Philp, K. Boesze-Battaglia, and D. S. Williams, "The cell biology of the retinal pigment epithelium," *Prog. Retinal Eye Res.*, vol. 78, Jan. 2020, Art. no. 100846.

[68] O. Bernard et al., "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?" *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2514–2525, Nov. 2018.

[69] B. Seo, D. Mariano, J. Beckfield, V. Madenur, Y. Hu, T. Reina, M. Bobar, M. H. Nguyen, and I. Altintas, "Cardiac MRI image segmentation for left ventricle and right ventricle using deep learning," 2019, *arXiv:1909.08028*.

[70] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3523–3542, Jul. 2022.

[71] X.-X. Yin, L. Sun, Y. Fu, R. Lu, and Y. Zhang, "U-Net-based medical image segmentation," *J. Healthcare Eng.*, vol. 2022, pp. 1–16, Apr. 2022, doi: 10.1155/2022/4189781.

[72] R. Belenkaya et al., "Extending the OMOP common data model and standardized vocabularies to support observational cancer research," *JCO Clin. Cancer Informat.*, vol. 5, pp. 12–20, Dec. 2021.

[73] R. Vashisht, K. Jung, A. Schuler, J. M. Banda, R. W. Park, S. Jin, L. Li, J. T. Dudley, K. W. Johnson, and M. M. Shervey, "Association of hemoglobin A_{1c} levels with use of sulfonylureas, dipeptidyl peptidase 4 inhibitors, and thiazolidinediones in patients with type 2 diabetes treated with metformin: Analysis from the observational health data sciences and informatics initiative," *JAMA Netw. Open*, vol. 1, no. 4, 2018, Art. no. e181755.

[74] M. Nazar, M. M. Alam, E. Yafi, and M. M. Su'ud, "A systematic review of human–computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques," *IEEE Access*, vol. 9, pp. 153316–153348, 2021.

[75] N. Mehdiyev and P. Fettke, "Explainable artificial intelligence for process mining: A general overview and application of a novel local explanation approach for predictive process monitoring," in *Interpretable Artificial Intelligence: A Perspective of Granular Computing*. Springer, 2021, pp. 1–28.

[76] M. Stierle, J. Brunk, S. Weinzierl, S. Zilker, M. Matzner, and J. Becker, "Bringing light into the darkness—A systematic literature review on explainable predictive business process monitoring techniques," 2021.

[77] M. S. Park, H. Son, C. Hyun, and H. J. Hwang, "Explainability of machine learning models for bankruptcy prediction," *IEEE Access*, vol. 9, pp. 124887–124899, 2021, doi: 10.1109/ACCESS.2021.3110270.

[78] M. Dastani, E. Gerding, C. M. Jonker, T. Norman, S. Stein, and V. Yazdanpanah. (2021). *Responsibility Research for Trustworthy Autonomous Systems (Blue Sky Ideas Track)*. [Online]. Available: https://underline.io/lecture/15188-responsibility-research-for-trustworthy-autonomous-systems-(blue-sky-ideas-track)

[79] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, "Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions," 2021, *arXiv:2112.11561*.

[80] F. Zerka, V. Urovi, A. Vaidyanathan, S. Barakat, R. T. H. Leijenaar, S. Walsh, H. Gabrani-Juma, B. Miraglio, H. C. Woodruff, M. Dumontier, and P. Lambin, "Blockchain for privacy preserving and trustworthy distributed machine learning in multicentric medical imaging (C-DistriM)," *IEEE Access*, vol. 8, pp. 183939–183951, 2020, doi: 10.1109/ACCESS.2020.3029445.

[81] É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "Explainability of deep vision-based autonomous driving systems: Review and challenges," 2021, *arXiv:2101.05307*.

[82] Q. Xu, Z. Yang, Y. Jiang, X. Cao, Y. Yao, and Q. Huang, "Not all samples are trustworthy: Towards deep robust SVP prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3154–3169, Jun. 2022, doi: 10.1109/TPAMI.2020.3047817.

[83] N. Banerjee, T. Giannetsos, E. Panaousis, and C. C. Took, "Unsupervised learning for trustworthy IoT," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, Jul. 2018, pp. 1–8.

[84] H. He, J. Gray, A. Cangelosi, Q. Meng, T. M. McGinnity, and J. Mehnen, "The challenges and opportunities of artificial intelligence for trustworthy robots and autonomous systems," in *Proc. 3rd Int. Conf. Intell. Robotic Control Eng. (IRCE)*, Aug. 2020, pp. 68–74.

[85] P. Marcillo, Á. L. V. Caraguay, and M. Hernández-Álvarez, "A systematic literature review of learning-based traffic accident prediction models based on heterogeneous sources," *Appl. Sci.*, vol. 12, no. 9, p. 4529, Apr. 2022, doi: 10.3390/app12094529.

[86] O. Pribyl, R. Blokpoel, and M. Matowicki, "Addressing EU climate targets: Reducing CO_2 emissions using cooperative and automated vehicles," *Transp. Res. D, Transp. Environ.*, vol. 86, Sep. 2020, Art. no. 102437.

[87] S. Martínez-Fernández, X. Franch, A. Jedlitschka, M. Oriol, and A. Trendowicz, "Developing and operating artificial intelligence models in trustworthy autonomous systems," 2020, *arXiv:2003.05434*.

[88] S. Pendleton, H. Andersen, X. Du, X. Shen, M. Meghjani, Y. Eng, D. Rus, and M. Ang, "Perception, planning, control, and coordination for autonomous vehicles," *Machines*, vol. 5, no. 1, p. 6, Feb. 2017, doi: 10.3390/machines5010006.

[89] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," 2019, *arXiv:1910.07738*.

[90] D. Omeiza, S. Anjomshoae, K. Kollnig, O.-M. Camburu, K. Främling, and L. Kunze, "Towards explainable and trustworthy autonomous physical systems," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2021, pp. 1–3, doi: 10.1145/3411763.3441338.

[91] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2921–2929.

[92] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," 2016, *arXiv:1603.08507*.

[93] M. Bojarski, A. Choromanska, K. Choromanski, B. Firner, L. Jackel, U. Müller, and K. Zieba, "VisualBackProp: Efficient visualization of CNNs," 2016, *arXiv:1611.05418*.

[94] M. Hofmarcher, T. Unterthiner, J. Arjona-Medina, G. Klambauer, S. Hochreiter, and B. Nessler, "Visual scene understanding for autonomous driving using semantic segmentation," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019, pp. 285–296.

[95] W. Zeng, W. Luo, S. Suo, A. Sadat, B. Yang, S. Casas, and R. Urtasun, "End-to-end interpretable neural motion planner," 2021, *arXiv:2101.06679*.

[96] Y. Xu, X. Yang, L. Gong, H.-C. Lin, T.-Y. Wu, Y. Li, and N. Vasconcelos, "Explainable object-induced action decision for autonomous vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9520–9529.

[97] J. Kim, S. Moon, A. Rohrbach, T. Darrell, and J. Canny, "Advisable learning for self-driving vehicles by internalizing observation-to-action rules," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9658–9667.

[98] D. Omeiza, H. Webb, M. Jirotka, and L. Kunze, "Explanations in autonomous driving: A survey," 2021, *arXiv:2103.05154*.

[99] S. V. Albrecht, C. Brewitt, J. Wilhelm, B. Gyevnar, F. Eiras, M. Dobre, and S. Ramamoorthy, "Interpretable goal-based prediction and planning for autonomous driving," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 1043–1049, doi: 10.1109/ICRA48506.2021.9560849.

[100] J. P. Hanna, A. Rahman, E. Fosong, F. Eiras, M. Dobre, J. Redford, S. Ramamoorthy, and S. V. Albrecht, "Interpretable goal recognition in the presence of occluded factors for autonomous vehicles," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 7044–7051, doi: 10.1109/iros51168.2021.9635903.

[101] H. Krasowski, X. Wang, and M. Althoff, "Safe reinforcement learning for autonomous lane changing using set-based prediction," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2020, pp. 1–7.

[102] U. Ehsan and M. O. Riedl, "Human-centered explainable AI: Towards a reflective sociotechnical approach," in *Proc. Int. Conf. Hum.-Comput. Interact.* Cham, Switzerland: Springer, 2020, pp. 449–466.

[103] U. Ehsan and M. O. Riedl, "Human-centered explainable AI: Towards a reflective sociotechnical approach," 2020, *arXiv:2002.01092*.

VINAY CHAMOLA (Senior Member, IEEE) received the B.E. degree in electrical and electronics engineering and the master's degree in communication engineering from the Birla Institute of Technology and Science (BITS), Pilani, India, in 2010 and 2013, respectively, and the Ph.D. degree in electrical and computer engineering from the National University of Singapore, Singapore, in 2016. In 2015, he was a Visiting Researcher with the Autonomous Networks Research Group (ANRG), University of Southern California, Los Angeles, CA, USA. He was a Postdoctoral Research Fellow with the National University of Singapore. He is currently an Assistant Professor with the Department of Electrical and Electronics Engineering, BITS-Pilani, where he heads the Internet of Things Research Group/Laboratory. His research interests include the IoT security, blockchain, UAVs, VANETs, 5G, and healthcare. He is a Fellow of IET. He is listed among the World's Top 2% Scientists identified by Stanford University. He is the Co-Founder and the President of Medsupervision Pvt., Ltd., a healthcare startup. He serves as the Co-Chair for various reputed workshops, such as IEEE Globecom Workshop 2021, IEEE INFOCOM 2022 Workshop, IEEE ANTS 2021, and IEEE ICIAfS 2021. He also serves as an Area Editor for the *Ad Hoc Networks* journal, Elsevier, and the *IEEE Internet of Things Magazine*, and an Associate Editor for the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE NETWORKING LETTERS, *IEEE Consumer Electronics Magazine*, *IET Quantum Communications*, *IET Networks*, and several other journals.

VIKAS HASSIJA received the B.Tech. degree from Maharshi Dayanand University, Rohtak, India, in 2010, the M.S. degree in telecommunications and software engineering from the Birla Institute of Technology and Science (BITS), Pilani, India, in 2014, and the Ph.D. degree in the IoT security and blockchain from the Jaypee Institute of Information and Technology (JIIT), Noida. He was an Assistant Professor with JIIT for four years. He holds a postdoctoral position with the National University of Singapore (NUS). He has eight years of industry experience and has worked with various telecommunication companies, such as Tech Mahindra and Accenture. His research interests include the IoT security, network security, blockchain, and distributed computing.

A RAZIA SULTHANA received the B.Tech. degree in information technology and the M.E. degree in computer science engineering from Anna University, Tamil Nadu, India, in 2007, the M.B.A. degree in systems from Madras University, Tamil Nadu, in 2009, the M.E. degree in computer science engineering from Anna University, in 2011, and the Ph.D. degree in computer science from the SRM Institute of Science and Technology, Tamil Nadu. She is an academic and an active researcher and has over 13 years of experience in delivering computer science and interdisciplinary courses for undergraduate and postgraduate students. Her international experience includes teaching students in India and Dubai. She has been working in the areas of computational intelligence, artificial intelligence, machine learning, and neural computing and is currently trying to interconnect her research intelligence with areas in edge computing and modeling technologies. She has published over 30 research papers in journals and conferences, published three patents, and owns an IBM AI explorer and mastery badge. She received the gold medal for her M.E. degree.

DEBSHISHU GHOSH is currently pursuing the B.Tech. degree in computer science engineering with the Jaypee Institute of Information Technology, Noida. He is a Data Engineering Intern at Krishify. He has worked on some projects in machine learning and is actively involved in the usage and applications of machine learning. His research interests include machine learning and deep learning applications in language processing and the healthcare domain.

DIVYANSH DHINGRA is currently pursuing the B.Tech. degree in computer science engineering with the Jaypee Institute of Information Technology, Noida. He is a Software Engineer Intern with Thales Group. His research interests include applications of machine learning and deep learning especially in the healthcare domain.

BIPLAB SIKDAR (Senior Member, IEEE) received the B.Tech. degree in electronics and communication engineering from North Eastern Hill University, Shillong, India, in 1996, the M.Tech. degree in electrical engineering from the Indian Institute of Technology Kanpur, Kanpur, India, in 1998, and the Ph.D. degree in electrical engineering from the Rensselaer Polytechnic Institute, Troy, NY, USA, in 2001. He was a Faculty Member with the Rensselaer Polytechnic Institute, from 2001 to 2013, an Assistant Professor, and an Associate Professor. He is currently a Professor and the Head of the Department of Electrical and Computer Engineering at the National University of Singapore, Singapore. His current research interests include wireless networks and security for the Internet of Things and cyber-physical systems. He served as an Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE INTERNET OF THINGS JOURNAL, and IEEE OPEN JOURNAL OF VEHICULAR TECHNOLOGY.