Ubiquitous Technology Journal
Vol. 1, No. 2, June 2025, pp. 46~60
ISSN: 3079-5273, DOI: 10.71346/utj.v1i2.21

Reasoning About Responsibility in Autonomous Systems: Navigating the Challenges and Charting Future Directions

Usman Tariq [1], Irfan Ahmed [2]
[1]Department of Management Information Systems, College of Business Administration, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia
[2]Department of Computer Science, College of Engineering, Virginia Commonwealth University, Richmond, VA, USA

Article Info

Article history:

Received March 05, 2025
Revised April 30, 2025
Accepted May 12, 2025

Keywords:

Autonomous Systems
Responsibility
Explainable AI
Ethical Principles
Hybrid Reasoning

ABSTRACT

As autonomous systems gain prominence in sectors such as transportation, healthcare, and finance, the challenge of assigning responsibility for their actions has become increasingly critical. Existing legal, ethical, and technical frameworks often fall short in addressing the unique characteristics of these systems, which include opaque decision-making processes, emergent behavior, distributed control, and learning from biased data. This paper investigates the core challenges involved in reasoning about responsibility within autonomous systems by focusing on issues such as the black-box problem, the unpredictability of outcomes, the complexity of multi-agent environments, and the evolving role of human oversight. It reviews and analyzes a range of potential solutions, including explainable AI techniques, formal specification & verification methods, agent-based simulations, ethics-oriented design principles, and hybrid reasoning models that combine symbolic & sub-symbolic approaches. By connecting these methods to real-world domains and incidents, the paper offers a structured understanding of how responsibility can be clarified and embedded into the design & governance of autonomous systems. This research contributes novel analytical perspectives and practical pathways that can support the more accountable deployment of AI technologies while laying the groundwork for future interdisciplinary probe into responsible autonomy.

Corresponding Author:

Usman Tariq
Department of Management Information Systems, College of Business Administration, Prince Sattam Bin Abdulaziz University

11942, Al-Kharj, Al-Riyadh, Saudi Arabia
Email: u.tariq@psau.edu.sa

1. INTRODUCTION

Autonomous systems have become increasingly integrated into modern life, with applications ranging from self-driving vehicles and robotic assistants to algorithmic decision-makers in financial services and healthcare. Their growing prevalence is not simply a reflection of technological progress but a signal of society's increasing reliance on computational agents that perform tasks with minimal human supervision [1]. As these systems acquire more sophisticated capabilities, they are being trusted to make decisions in situations characterized by uncertainty, limited data, and evolving environmental conditions [2]. The greater their autonomy, the more critical the implications when decisions lead to unintended or adverse outcomes. This shift in operational control raises pressing questions about responsibility, especially when the causality behind a system's actions is difficult to trace or explain.

The issue of responsibility becomes particularly complex when autonomous systems operate in settings where outcomes affect human welfare, safety, or rights [3]. Traditional models of accountability often presume direct human oversight and intentional agency. In contrast, systems driven by machine learning or adaptive algorithms do not always conform to predictable patterns of behavior. They may exhibit emergent properties or make decisions that even their designers cannot fully anticipate or justify [4]. This disconnect creates a conceptual gap between system behavior and human accountability which makes it difficult to determine who should bear the consequences when things go wrong. The challenge lies not only in understanding the mechanisms underlying these decisions but also in framing them within moral, legal, and technical structures that are equipped to manage such complexity [5].

Existing approaches to assigning responsibility in technological contexts are often grounded in frameworks designed for deterministic or rule-based systems. These methods fall short when applied to autonomous agents that exhibit learning capabilities and operate in partially observable environments. One major limitation is the lack of formal models that can handle the distribution of decision authority across multiple agents and human stakeholders [6]. Likewise, the opacity of many current AI architectures further complicates this task, as system outputs are not always interpretable even by those who designed them [7]. The presence of non-deterministic behavior and evolving decision policies introduces a moving target for any model of responsibility. Without an adequate foundation to account for these variables, attributions of blame or liability risk being arbitrary or ineffective.

This paper is driven by the need to better understand and systematize reasoning about responsibility in autonomous systems. It aims to tackle the core problem of assigning responsibility in environments where actions are initiated or shaped by agents that may not have human-like intent or awareness. The research critically examines the interplay between technical features—such as algorithmic learning, system opacity, and distributed control—and normative principles that guide ethical and legal accountability. It also scrutinizes the structural deficiencies in current responsibility models and proposes directions for enhancing their robustness. This inquiry is vital to developing systems that not only perform reliably but are also aligned with societal expectations for transparency, fairness, and justice.

The paper addresses the following research questions.

1. What are the principal challenges in establishing responsibility frameworks for highly autonomous systems operating under uncertainty and minimal human input?
2. In what ways can techniques from explainable artificial intelligence contribute to clarity in responsibility attributions across stakeholders?
3. How can ethical and legal principles be systematically integrated into the technical design of autonomous agents to enable responsibility-aware behavior?

These questions define the scope of this study and anchor its theoretical and methodological contributions.

The contributions of this paper are fourfold.

a) It introduces a formal conceptual model that integrates causal reasoning with normative assessment to evaluate responsibility in autonomous systems.
b) It presents a taxonomy of responsibility failure modes based on system design attributes and operational contexts.
c) It evaluates the efficacy of selected explainability techniques in aiding responsibility judgments through empirical simulation.
d) Finally, it proposes a set of design principles for embedding responsibility considerations during system development.

Each of these contributions fills a critical gap in current literature and provides a scaffold for future inquiry in this domain.

To ensure a rigorous and systematic selection of relevant literature, a PRISMA-based methodology was employed. An initial pool of 4827 records was gathered from scholarly databases including ‘Web of Science SCIE’, ‘Google Scholar’, and ‘Scopus’. Following a structured screening process, 2549 records were excluded due to various reasons such as duplication and irrelevance to the research scope. Titles and abstracts of 672 studies were subsequently reviewed for eligibility, with 259 citations excluded due to insufficient information. Ultimately, 75 studies were selected for in-depth analysis, which encompass key themes such as autonomous systems, responsibility, explainable AI, formal methods, agent-based modeling, ethical principles, legal frameworks, hybrid reasoning, auditing, and certification. This transparent and methodical approach strengthens the comprehensiveness and credibility of the review by grounding it in a robust scholarly foundation.

Thus, the significance of this research lies in its interdisciplinary synthesis and forward-looking perspective. Rather than treating responsibility as an afterthought to system design, this paper positions it as a core design parameter. The projected models and frameworks are not limited to any specific domain and are

applicable across a wide range of autonomous systems, from military drones to healthcare diagnostics. By establishing a foundation for rigorous reasoning about responsibility, this research investigation contributes to ongoing efforts in making autonomous systems both safe and trustworthy. It offers a structured path forward for engineers, ethicists, and policymakers aiming to align technological progress with human values and social norms.

The paper progresses from establishing the Foundations of Responsibility in Autonomous Systems, exploring philosophical, ethical, legal, and technical perspectives, and discussing key concepts like agency, intentionality, causality, and foreseeability. It identifies Challenges in Reasoning About Responsibility, such as the black box problem, unforeseen circumstances, emergent behavior, distributed agency, data bias, temporal aspects, and human oversight. Then, it explores Opportunities and Approaches for Reasoning About Responsibility, including explainable AI, formal methods, agent-based modeling, ethical principles, hybrid approaches, and regulatory frameworks. Following this, the paper presents Case Studies and Applications, analyzing responsibility in specific domains like autonomous vehicles, healthcare robotics, and AI in financial decision-making, and examining real-world incidents. Finally, the paper Synthesizes Findings, Implications, and Future Directions by summarizing key challenges and opportunities, implications for design, development, and deployment, and addressing the research questions, ultimately concluding with a call for interdisciplinary collaboration and proactive responsibility engineering.

2. FOUNDATIONS OF RESPONSBILITY IN AUTONOMOUS SYSTEMS

2.1 Exploring Philosophical, Ethical, Legal, and Technical Perspectives

Responsibility is a concept that encompasses a rich spectrum of interpretations depending on disciplinary focus and context. In philosophy, it is often associated with the capacity to act with moral awareness and to be held accountable for the consequences of one's actions. Moral responsibility is commonly rooted in agency and intentionality, notions which become problematic when applied to autonomous systems that operate based on algorithmic decision-making rather than conscious deliberation [8-9]. The absence of consciousness and volition challenges conventional interpretations of blameworthiness and ethical accountability. Scholars have questioned whether systems without desires or beliefs can truly be moral agents or if responsibility must instead be redirected toward human stakeholders who design, deploy, or benefit from these systems [10]. From an ethical perspective, the issue is compounded when such systems make decisions that significantly impact human lives without any clear path for moral recourse or redress [11].

In ethical theory, responsibility for autonomous systems has been approached from various angles. Discussions often focus on the degree to which systems can be embedded with ethical principles or behavioral constraints that reflect human values. The challenge lies in developing mechanisms that allow these systems to evaluate potential harms or benefits while maintaining the adaptability required for complex tasks. The notion of ethical accountability becomes more nuanced in multi-agent contexts where collective behavior emerges from decentralized decision-making [12]. This raises further questions about the diffusion of responsibility across interacting agents. Meanwhile, in legal contexts, the question of responsibility centers on liability, culpability, and due diligence [13]. Traditional legal frameworks are anchored in principles such as negligence, product liability, and foreseeability. These constructs presuppose a clear actor or institution that can be identified and held to account when harm occurs. When autonomous systems make decisions that lead to harmful outcomes, these frameworks strain under the weight of ambiguity which prompts discussions on the development of new legal categories or doctrines specifically tailored to accommodate autonomous agents [14].

Technical perspectives on responsibility have gained traction in the effort to design systems that are not only functional but also auditable and interpretable. Concepts such as algorithmic accountability and auditability aim to support post hoc analysis of decision-making processes in systems powered by machine learning or reinforcement learning algorithms. Unlike deterministic systems, learning-based models can evolve over time which leads to behavior that is difficult to predict or replicate. This adaptability presents a significant challenge to traditional verification and validation techniques. Researchers have proposed design principles, as illustrated in Taxonomy (i.e., Table 1), that embed responsibility into the system architecture by incorporating explainability features, decision provenance, and structured feedback mechanisms. These features serve to narrow the accountability gap between human stakeholders and the systems they engineer. Yet, the integration of these technical constructs with ethical and legal norms remains an open area of investigation.

Table 1. Taxonomy of Challenges in Reasoning About Responsibility in Autonomous Systems

Taxonomy Category	Challenge Description	Key Attributes	Implications for Responsibility	Potential Mitigation Strategies
Technical Complexity				
Opacity of AI Models [15-16]	Many AI models, especially deep neural networks, have internal structures that are difficult to interpret.	Complex layers, non-linear computations, feature interactions	Hinders tracing decision pathways, complicates accountability, limits auditability.	Explainable AI (XAI) techniques, rule-based models
Unforeseen Circumstances & Emergent Behaviour [17]	Autonomous systems in dynamic environments may encounter novel situations, leading to unpredictable behaviour.	Complex interactions, environmental variability, adaptive learning	Challenges traditional accountability models, makes foreseeability difficult, diffuses responsibility.	Agent-based modelling, robust testing, continuous monitoring
Distributed Agency [18]	Multi-agent systems involve multiple autonomous entities, making it hard to isolate the source of an outcome.	Decentralized control, collective objectives, coordination protocols	Diffuses responsibility across agents, complicates individual vs. system-level accountability.	Formal methods, clear role definitions, accountability frameworks
Data-Related Issues				
Data Bias [19-20]	Training data may contain biases that lead to skewed or discriminatory outcomes.	Historical inequities, skewed sampling, feature selection bias	Results in unfair or discriminatory outcomes, obscures root causes of harm.	Fairness constraints, data audits, diverse datasets
Temporal Dynamics				
Temporal Aspects [21-22]	Autonomous systems may evolve over time, making it difficult to assign responsibility for actions taken after deployment.	Continuous learning, adaptation, shifting behaviour	Creates a moving target for accountability, complicates retrospective and prospective responsibility.	Versioning, logging, adaptive governance
Human-System Interaction				
Human Oversight Limitations [23-24]	Human oversight of autonomous systems has inherent limitations in high-speed or complex situations.	Limited attention, information overload, intervention challenges	Raises questions about the role and responsibility of human operators.	Clear guidelines, effective human-machine interfaces, training

2.2 Levels of Autonomy and Their Implications for Responsibility

The attribution of responsibility becomes even more complex when considering varying degrees of autonomy. Autonomy in artificial systems is typically modeled on a continuum ranging from fully manual to fully autonomous operation. At lower levels of autonomy, human operators maintain direct control over system behavior and decision-making remains traceable to user input. As autonomy increases, decision authority shifts toward the system itself, often resulting in less human oversight and intervention. This redistribution introduces ambiguities in accountability, especially when adverse events arise. Designers and developers must anticipate the consequences of delegating control [25], while operators must contend with the unpredictability of system behavior [26]. As systems reach higher autonomy levels, the involvement of human decision-makers becomes increasingly indirect which exacerbates what scholars refer to as the

responsibility gap. This gap represents the difficulty in pinpointing a morally or legally appropriate locus of accountability in cases where system actions deviate from intended outcomes.

2.3 Key Concepts: Agency, Intentionality, Causality, Foreseeability in the Context of Autonomous Systems

Agency, intentionality, causality, and foreseeability are foundational concepts that further complicate responsibility in autonomous systems.

a) The idea of agency typically implies the capacity to act with intention toward a goal, yet in autonomous systems agency is an engineered abstraction driven by algorithmic policies and reward structures. Some scholars argue that this form of artificial agency lacks the normative significance of human agency because it does not involve consciousness or volitional intent [27].
b) Intentionality [28], when applied to machines, is therefore interpreted functionally rather than phenomenologically.
c) Causality [29] plays a central role in tracing responsibility, especially when an undesired outcome must be linked back through the system's operational chain to a decision node. Establishing a causal link between a system's action and a harmful result is critical for attributing responsibility. Yet the probabilistic nature of many AI algorithms complicates this traceability.
d) Foreseeability [29] refers to the capacity to predict potential harms before deployment. The responsibility to foresee adverse consequences typically falls on system designers, but the adaptive behavior of modern AI challenges the scope and accuracy of such foresight.

2.4 Existing Frameworks for Responsibility and Their Applicability to Autonomous Systems

In recent years, scholars and practitioners have evaluated numerous frameworks for responsibility attribution from ethics, law, and human-computer interaction [1][31]. Legal doctrines such as strict liability and negligence have been analyzed in light of autonomous systems that act independently of direct human control. Ethical theories of moral responsibility [9] have been extended to include system designers and organizations as collective moral agents. Within human-computer interaction, models emphasizing transparency, user trust, and system explainability aim to support responsible human engagement with automated systems. Each of these frameworks brings valuable insights, yet they also display limitations when applied to systems that exhibit autonomy, adaptivity, and non-linear learning. For instance, strict liability assumes a static system behavior profile, which is rarely the case in real-world deployments [32]. Similarly, transparency models in HCI often focus on user-facing aspects rather than internal system processes that are critical for deeper accountability [33]. Bridging these gaps requires an integrative approach that combines normative principles with system-level technical rigor.

2.5 Challenges in Reasoning About Responsibility

The Black Box Problem: Opacity and Lack of Explainability in Complex AI Models

One of the most persistent challenges in reasoning about responsibility in AI systems lies in the black box nature of many modern machine learning models [34-36]. Deep neural networks, in particular, are characterized by internal structures that defy straightforward interpretation. Although they are capable of identifying complex patterns and making accurate predictions, their decision-making processes are often opaque even to their creators. This opacity impedes efforts to identify the rationale behind specific actions or predictions, particularly when outcomes have ethical or legal implications. The absence of a clear explanatory pathway limits our ability to assign responsibility when an AI system causes harm, and it complicates efforts to audit or refine system behavior post hoc. Without reliable interpretability, AI systems risk being used in high-stakes domains without the safeguards required for accountability.

Likewise, the lack of explainability [37] directly affects our capacity to trace causality in AI decision-making. When a system fails or produces biased results, the underlying factors are frequently buried within a maze of hidden layers and nonlinear computations. This presents a major obstacle for identifying whether an error was due to flawed training data, inappropriate architectural choices, or external context not accounted for during development. In the absence of this clarity, responsibility may be incorrectly assigned or, worse, evaded altogether. The inability to identify actionable causes also hinders remediation efforts, as developers may not know which components need to be adjusted or retrained. In such scenarios, the lack of transparency not only frustrates ethical inquiry but also undermines public trust in AI systems designed for social or institutional use.

2.6 The Problem of Unforeseen Circumstances and Emergent Behavior

Another major challenge arises from the behavior of AI systems in real-world environments that are neither static nor fully predictable [2]. These systems are increasingly deployed in contexts where inputs evolve and interact in complex ways. AI trained under constrained conditions may encounter inputs or combinations of stimuli it has never seen before which leads to decisions that were not explicitly foreseen by developers. This capacity for adaptation introduces a significant degree of unpredictability into system behavior. Emergent behavior [12] further complicates the issue, as complex responses can arise from interactions among relatively simple subcomponents. These emergent effects are not always intuitive or traceable to a single element of system design, thereby weakening the prospects for assigning individual or even collective responsibility in a coherent manner.

The implications of such unpredictability are profound when it comes to accountability [1][38]. Traditional accountability models rely on the assumption that actions are traceable, and outcomes are foreseeable. In systems where new patterns of behavior emerge over time, this assumption no longer holds. Determining liability in such cases becomes a question of degree rather than clarity. For instance:

a) Did the developers anticipate sufficient variability?
b) Were operators equipped to intervene effectively?

These questions point to a significant structural gap in current frameworks, which are not always equipped to handle the fluid nature of emergent AI behavior. This gap becomes particularly problematic when systems deployed in critical settings such as healthcare or transportation encounter edge cases not captured during training.

2.7 Distributed Agency and Responsibility in Multi-Agent Systems

As AI systems evolve toward more collaborative and distributed architectures, the challenge of reasoning about responsibility is magnified [1-2][4][39]. Multi-agent systems involve multiple autonomous entities working toward individual or collective objectives, often coordinating through decentralized protocols. In such environments, it becomes exceedingly difficult to isolate the source of an outcome. Responsibility [3][40] becomes diffused across the agents and, by extension, the individuals and organizations who designed and deployed them. Each agent may make decisions based on local information without full knowledge of the global system state. This decentralized decision-making structure complicates attribution and raises fundamental questions about whether responsibility can meaningfully reside at the level of individual agents or must instead be assessed at the system level.

The distinction between individual and systemic responsibility is not always straightforward in these scenarios. While it may be tempting to view each agent as accountable for its own actions, the collective behavior of the system may produce outcomes that no single agent intended or predicted. This tension challenges current ethical and legal standards that often rely on direct causality and clearly defined actors. Furthermore, when systems include both human and artificial agents, the complexity of assigning roles and duties increases substantially. Questions emerge about the extent to which human supervisors should monitor agent interactions and whether oversight itself should be distributed across the system. These ambiguities demand new models that can represent and evaluate responsibility in complex sociotechnical ecosystems [41].

2.8 The Influence of Data and Bias on Responsibility Attribution

Another deeply entrenched issue stems from the role of data in shaping AI behavior and influencing accountability [42-43]. Machine learning systems are profoundly dependent on the quality, diversity, and structure of the data they are trained on. Biases [44-45] embedded within training data can produce skewed outcomes that systematically disadvantage certain groups or individuals. When such biased behavior is revealed, assigning responsibility becomes a contested exercise. For instance:

a) Was it the fault in the original data collection process?
b) Did the developers fail to apply appropriate debiasing techniques? Or
c) Did the deployers misjudge the applicability of the model in a given context?

Each of these points represents a different locus of responsibility, and disentangling them is neither simple nor trivial.

Bias in data can also obscure the root causes of harmful or discriminatory outcomes. If the data itself reflects historical inequalities or social prejudices, then the model may inadvertently replicate or exacerbate those patterns. This introduces epistemic opacity, where even the causes of bias remain hidden or misunderstood. In such cases, the standard tools of responsibility attribution may fall short, as they presuppose a degree of clarity that biased models often do not provide. These complications underscore the importance of transparency and documentation not only in model architecture but throughout the entire data

pipeline. Yet even with comprehensive audits, the influence of bias may persist in subtle and unexpected ways which makes accountability a continuous and evolving challenge.

2.9 Temporal Aspects of Responsibility: Responsibility for Past, Present, and Future Actions

Temporal aspects of responsibility introduce another layer of complexity in ethical AI design [46]. AI systems are often deployed with the capacity to learn and adapt over time. As they interact with new data or user environments, their behavior may shift in ways not anticipated at the time of deployment. This raises the question of how to assign responsibility for actions taken by systems that no longer resemble their original configurations. Developers may argue that they cannot be held accountable for outcomes resulting from post-deployment learning, while operators may lack the technical capacity to fully understand or control the system's evolution [15]. This creates a moving target for accountability, particularly in systems that operate autonomously over long periods.

Looking forward, responsibility must also be considered for potential future harms. AI systems designed today may be integrated into environments that pose risks not yet evident. Regulatory and ethical frameworks must account for this forward-looking dimension. Assigning prospective responsibility—anticipating and planning for potential failures—is a necessary component of responsible innovation. Yet doing so is fraught with difficulty given the uncertainty of future contexts and interactions. Risk assessments [47] must be iterative and adaptive, incorporating new knowledge as it emerges. Without this capacity, responsibility becomes retrospective and reactive, rather than proactive and preventive, undermining the goal of ethical foresight in AI governance.

2.10 The Challenge of Human Oversight and Intervention: Defining the Role and Responsibility of Human Operators

Human oversight continues to be framed as a critical safeguard against AI errors, but defining its role presents significant ethical and practical challenges. Oversight is often touted as a mechanism to preserve human control and judgment, yet its effectiveness depends on context, timing, and the design of the system itself. In many high-speed or high-volume applications, the window for meaningful human intervention is limited or nonexistent. Even when intervention is possible, human operators may lack sufficient information or understanding to make informed decisions. This raises the question of whether oversight can serve as a reliable anchor for responsibility or if it simply creates the illusion of control.

Scenarios where human intervention contributes to failure further complicate responsibility attribution. A human may override an automated recommendation based on flawed intuition or incomplete data, leading to an undesirable outcome. In such cases, responsibility is shared, yet not always in equal measure. Defining appropriate divisions of responsibility between human and machine requires models that account for both capacity and context. These models must specify not only what actions are possible but what actions are expected or reasonable under particular conditions. Failing to do so risks placing undue blame on individuals or letting systemic flaws go unaddressed. Robust ethical design [9] must therefore include clearly defined parameters for oversight and response, embedded at both operational and organizational levels.

3. RESULTS AND DISCUSSION (10 PT) OPPORTUNITIES AND APPROACHES FOR REASONING ABOUT RESPONSIBILITY

3.1 Explainable AI (XAI) for Responsibility Attribution

Explainable AI techniques represent a promising response to the challenge of opaque decision-making in autonomous systems. Methods such as SHAP and LIME [35] provide post hoc explanations by attributing output behavior to input features while attention mechanisms in neural networks offer insights into internal focus during prediction. Rule-based models [48] and surrogate explanations [49] add another layer of interpretability by generating logical statements or simplified models that approximate original system behavior. These techniques can support responsibility attribution by revealing which factors influenced a decision and how these factors interacted to yield an outcome. Increased transparency aids not only developers in debugging but also allows stakeholders such as end users and legal professionals to assess fairness and appropriateness. Tailoring explanations for different audiences remains a research priority as the technical depth required by engineers may not align with the interpretive needs of regulatory bodies. Although current methods [50] offer useful tools [51], limitations persist in generalizing explanations across domains and preserving fidelity to complex model behavior. In the stated context, future developments & deployments must improve explanation reliability, contextual relevance, and integration with formal verification tools for enhanced responsibility reasoning.

3.2 Formal Methods and Logical Frameworks for Specifying and Verifying Responsibility

Formal methods [6][52] offer structured avenues for specifying norms and obligations in autonomous systems. Approaches using temporal logic [53] allow designers to specify system behavior over time while deontic logic [54] provides a means of expressing duties and permissions. Such logical tools can formalize notions of expected conduct and support verification of compliance with defined responsibilities. This makes it possible to detect violations systematically and enforce accountability through automated reasoning mechanisms. Formal methods also allow for rigorous proof of behavioral properties which is crucial when autonomous systems are deployed in safety-critical domains. Despite these advantages, the translation of ethical and legal norms into formal specifications is inherently complex. Ethical guidelines are often context-sensitive and ambiguous, which makes them difficult to encode with precision. Scaling these methods to account for interactions among numerous agents with heterogeneous goals further complicates the task. Advances in domain-specific logics and hybrid formalisms are needed to manage such complexity while maintaining tractability.

3.3 Agent-Based Modeling and Simulation for Exploring Responsibility Dynamics

Agent-based modeling [55] offers a powerful tool for studying responsibility allocation in systems composed of multiple autonomous actors. These simulations create controlled environments in which agents follow predefined rules and interact with one another that allow researchers to observe the resulting emergent behavior. Through this approach, it becomes possible to analyze how variations in communication protocols, control structures, or task assignments influence responsibility diffusion. Such models are useful for identifying failure points to predict blame propagation patterns and evaluate the effectiveness of accountability mechanisms before real-world deployment. They can simulate both intended outcomes and unanticipated effects to provide insight into the resilience of responsibility frameworks under stress. Despite their promise, the fidelity of agent-based simulations depends heavily on accurate representation of agent behavior and environmental complexity. Hereby, the validation of simulation outcomes remains a challenge, and the conclusions drawn must be interpreted within the boundaries of model assumptions. Nevertheless, simulations serve as valuable testbeds for prototyping and refining responsibility reasoning frameworks.

3.4 Incorporating Ethical Principles and Legal Frameworks into Autonomous System Design

Embedding ethical and legal considerations directly into the design process marks a shift toward proactive responsibility engineering. Frameworks such as value-sensitive design [56] and ethics by design [9][11][23][44][54] advocate for early integration of societal values, transparency mechanisms, and normative constraints. This involves translating abstract principles like fairness or accountability into concrete system-level rules and parameters. Techniques may include constraint programming to enforce non-discrimination or procedural safeguards to allow for human review and contestability. Legal compliance [15] can be encoded through rule sets derived from relevant regulatory texts or adapted standards. The principal difficulty lies in resolving tensions among competing ethical objectives and adjusting system behavior in response to evolving norms. Legal interpretations may differ across jurisdictions, requiring systems to support localization and adaptability. As a result, multi-disciplinary collaboration becomes essential for operationalizing these norms in a meaningful and consistent manner across system contexts.

3.5 Hybrid Approaches: Combining Symbolic and Sub-symbolic Reasoning for Robust Responsibility Assessment

Hybrid AI techniques [17] hold great promise for synthesizing robust and adaptable responsibility reasoning. Symbolic methods offer formal tools for encoding norms, rules, and structured knowledge, while sub-symbolic models excel in pattern recognition and adaptability to real-world data [57]. Combining these strengths allows for the creation of systems that can represent high-level concepts such as obligation or liability while also responding to novel inputs through learned associations. For instance, a symbolic layer may define ethical guidelines while a neural network layer manages environmental perception and tactical decisions. The integration of these components allows for reasoning that is both principled and context-aware. Herewith, the coordination between symbolic and sub-symbolic reasoning remains a technical challenge. Likewise, the mapping between representational formats and maintaining alignment between learned and encoded knowledge requires careful design. Thus, we reckon that the future advancements in neuro-symbolic architectures may offer solutions by introducing shared representational spaces or modular reasoning pipelines that maintain coherence across layers.

3.6 The Role of Auditing, Certification, and Regulatory Frameworks

Institutional mechanisms such as audits and certification provide structured means for enforcing accountability in AI development and deployment. Independent audits can examine whether systems meet transparency, fairness, and robustness standards, while certification can formalize compliance with industry's best practices [15][58-60] and legal requirements. These processes serve as external checks that encourage responsible innovation and build public trust. Regulatory frameworks [61-62] can further support accountability by establishing mandatory reporting, documentation requirements, and oversight protocols tailored to specific sectors. Creating such frameworks presents challenges, including defining measurable criteria, maintaining technological neutrality, and adapting to evolving system capabilities. Regulatory lag [63] is a persistent concern as lawmaking and standardization often trail technological progress. Therefore, adaptive governance models [17][64] that incorporate real-time feedback and participatory review are essential for keeping accountability mechanisms relevant and effective. Together, these institutional structures play a key role in fostering an ecosystem where responsibility is not only a design consideration but a sustained operational priority.

4. CASE STUDIES AND APPLICATIONS

4.1 Examining Responsibility in Specific Autonomous System Domains

Autonomous vehicles [10][39][47][58] represent a critical domain where reasoning about responsibility has gained urgent importance. These systems operate in real-time environments, making high-stakes decisions based on sensory data and learned driving policies. The primary stakeholders include vehicle manufacturers, AI developers, passengers, pedestrians, and traffic regulators. One of the most pressing concerns arises when the vehicle makes a split-second decision that results in harm. While traditional liability might fall on the manufacturer, the involvement of adaptive learning systems introduces ambiguity. Questions emerge around whether the system was properly trained, whether edge cases were sufficiently accounted for, and whether a human should have intervened. Approaches such as explainable AI [35][37][51] can assist post-incident analysis by clarifying how decisions were made, while formal verification methods [6] could contribute to preemptively ruling out certain unsafe behaviors through bounded guarantees.

In the domain of healthcare robotics [65], responsibility is interwoven with ethical imperatives related to patient safety, consent, and decision transparency. Systems such as surgical robots and AI diagnostic tools often assist or replace human professionals in clinical settings. The stakeholders include patients, medical practitioners, system developers, hospital administrators, and regulators. Responsibility challenges intensify when outcomes diverge from expectations, particularly in diagnosis or surgical execution. The allocation of blame becomes complex when the robot executes a function based on sensor input and a learned protocol. Regulatory guidelines [66] often lag behind these advances, and the absence of clear explainability can obstruct both legal proceedings and patient redress. Ethical design strategies that incorporate normative constraints and simulate edge-case scenarios could improve responsibility clarity and reduce harm.

AI in financial decision-making [67-68], particularly in loan approval and credit scoring, introduces responsibility challenges related to fairness, transparency, and discrimination. Financial institutions, data providers, algorithm developers, and regulatory agencies are key actors. These systems use historical data to model risk and eligibility, which can embed and perpetuate biases. Applicants who are denied services based on opaque criteria often lack recourse due to the absence of meaningful explanations. In this context, SHAP values [69] and rule-based explainability [70] can offer insight into the influence of input features. Likewise, embedding fairness constraints during model training and incorporating value-sensitive design principles can prevent harm and make systems more accountable. These practices not only clarify responsibility but also support compliance with emerging regulatory frameworks related to algorithmic fairness.

4.2 Analyzing Real-World Incidents and the Challenges of Determining Responsibility

One notable incident occurred in 2018 involving an autonomous vehicle operated by a ride hailing service, which struck and killed a pedestrian during a test drive in Arizona [71]. The vehicle's sensors detected the pedestrian but failed to trigger an appropriate response. Investigations revealed that while the system identified the object, it misclassified its behavior and did not initiate braking. The human safety driver was distracted and did not intervene. Responsibility in this case spanned across multiple layers including the software's object classification algorithm, the system's decision fusion module, and the operator's oversight. Legal and ethical assessments faced difficulties in parsing which element failed most significantly and whether existing liability laws were equipped to deal with AI-induced harm. The incident highlighted the

need for continuous monitoring, clearer division of human and system roles, and more robust simulation testing prior to deployment.

In another case, a major health provider in the United States deployed an AI system to prioritize patients for intensive care. The model disproportionately assigned lower risk scores to Black patients due to biased historical data in its training set [72-73]. This led to delayed or denied care and raised critical concerns about systemic discrimination. Determining responsibility involved tracing the source of bias to the data collection and feature selection stages. Although developers may not have intentionally introduced bias, their failure to audit training data for representational equity made them complicit in the harmful outcomes. This case underscored the limitations of current development workflows that do not prioritize fairness and highlighted the importance of independent auditing and bias detection protocols as part of ethical AI development pipelines.

4.3 Illustrating the Application of Projected Reasoning Approaches through Concrete Examples

To illustrate the potential of explainable AI in responsibility attribution, consider a hypothetical extension of the ride hailing service provider's incident where the vehicle used SHAP-based explanations to log real-time decision contributions. In this case, the system could generate an explanation indicating that pedestrian detection had a high confidence score, but the trajectory prediction module assigned low threat probability due to poor environmental lighting. With this record, investigators could precisely identify the subsystem that failed and assess whether thresholds were appropriately calibrated. Legal authorities could use this information to assign responsibility to the module developers and recommend calibration updates for similar future deployments.

A second example can be drawn from healthcare robotics where formal methods are integrated into surgical planning software. Suppose a robotic assistant uses temporal logic specifications to verify each planned motion against a set of safety constraints. During a procedure, the robot halts an incision after detecting an unexpected tissue density that violates a predefined safety condition. The halt is logged along with the violated rule. In post-operative review, this log helps surgeons and system developers verify that the system followed ethical constraints and averted harm by design. This scenario demonstrates how formal verification not only prevents errors but also produces traceable evidence for responsibility reasoning and legal protection.

These case studies and hypothetical applications provide grounded evidence for the practical importance of responsibility reasoning in autonomous systems. They reveal both the depth of the challenges involved and the potential of current methodologies to address them. Whether through transparency-enhancing technologies or rule-based verification techniques, the emerging toolkit for responsible AI holds the promise of guiding safer and more ethically aligned system deployments in complex environments.

5. CONCLUSION

This research was motivated by the growing need to understand and assign responsibility in autonomous systems as they increasingly operate in high-stakes environments where human oversight is minimal, and outcomes are often difficult to predict. The central challenge addressed in this paper concerns the difficulty of reasoning about responsibility in systems that are opaque in design, prone to emergent behavior, structured around distributed agency, trained on potentially biased data, subject to temporal evolution, and reliant on varying degrees of human intervention. To address these challenges, the paper surveyed a range of promising approaches including explainable AI for traceable decision-making, formal methods for normative specification and verification, simulation-based modeling for examining responsibility propagation, ethics-embedded system design, hybrid AI architectures that combine rule-based and learned reasoning, and institutional mechanisms such as auditing and certification for external oversight. The research contributes novel insights into how these approaches map to specific responsibility challenges and highlights their potential for creating more transparent and accountable autonomous systems. It also provides a structured framework for understanding where current methods fall short and where future efforts should be concentrated. Looking ahead, the question of responsibility will remain central as AI systems become more autonomous and embedded in everyday life. Continued research is essential in developing tools that make decision pathways more intelligible to formalize ethical constraints that can be computationally implemented and establishing regulatory structures that are adaptable to technological change. A key takeaway is that responsibility must not be retrofitted after deployment but engineered as a foundational element of autonomous systems from the outset. Long-term progress in this field requires interdisciplinary collaboration that brings together insights from computer science, ethics, law, and the social sciences. As this

paper has shown, while the challenge is substantial, the pathway toward responsible and accountable autonomous systems is attainable through sustained theoretical and practical innovation.

5.1 Key Challenges and Opportunities Identified

The central challenges in reasoning about responsibility in autonomous systems arise from their technical complexity and sociotechnical entanglement. Opacity in AI models [74-75], particularly those based on deep learning, makes it difficult to trace decision pathways and identify accountable components. Unforeseen circumstances and emergent behaviors further complicate attribution, as outcomes may result from interactions beyond the scope of original design. Distributed agency in multi-agent systems introduces ambiguity in assigning responsibility across agents, especially when collective actions yield unintended effects. Data bias [44][73] remains a persistent concern as historical inequities encoded in training data can lead to discriminatory behavior. Temporal aspects [53] add complexity when systems evolve after deployment, raising questions about retrospective and prospective responsibility. Human oversight, while often suggested as a safeguard, faces practical and epistemic limitations in high-speed and high-volume applications.

To address these challenges, several promising approaches have emerged. Explainable AI [4][34][37][45][50-51] provides mechanisms for generating interpretable decision outputs to support causal tracing and stakeholder accountability. Formal methods [6][22][52] enable the specification and verification of behavioral constraints, making violations detectable by design. Agent-based modeling [17][55][64] offers a simulation framework to study how responsibility might propagate across agents under varying conditions. Ethical and legal principles [9][11][23][44][54][66] can be embedded during the design phase through frameworks like value-sensitive design. Hybrid approaches [57] that combine symbolic and sub-symbolic reasoning help balance rule-based transparency with data-driven adaptability. Institutional mechanisms such as auditing and certification [13] further enhance accountability by external evaluation and compliance enforcement. Each of these approaches offers a pathway for mitigating specific responsibility challenges. For instance, formal verification is well suited to address specification breaches, while explainable AI aids in interpreting opaque decision chains.

5.2 Implications for the Design, Development, and Deployment of Autonomous Systems

The findings from this study suggest that responsibility must be addressed as a core feature throughout the lifecycle of autonomous system development. Design principles must account for traceability, explainability, and ethical alignment from the earliest stages of conceptualization. Development methodologies should include continuous risk assessment, value-sensitive modeling, and integration of feedback from diverse stakeholders. At deployment, operational protocols must be prepared to support real-time logging, monitoring, and intervention when anomalies occur. These adjustments are not merely technical but institutional and require coordinated action from developers, regulators, legal practitioners, and users. Herewith:

a) Developers must prioritize auditability and fairness in system architecture.
b) Policymakers should establish adaptive regulatory standards that reflect the evolving capabilities of AI.
c) Legal professionals will need frameworks that accommodate non-human agency and distributed causality.
d) End-users should be equipped with understandable explanations and recourse mechanisms in case of system failure.

Without such multi-level engagement, responsibility may remain diffused and unaddressed in critical scenarios.

5.3 Addressing the Research Questions Posed in the Introduction

This research began with three key questions. First, what are the principal challenges in developing responsibility frameworks for highly autonomous systems? The analysis has shown that challenges arise from model opacity, unpredictable behavior, distributed decision-making, and temporal complexity. Second, how can explainable AI techniques contribute to transparent and accountable responsibility assignment? The study has demonstrated that XAI can clarify the internal logic of AI outputs, support post-hoc analysis, and provide stakeholder-specific justifications, thereby facilitating informed accountability. Third, what considerations are necessary for integrating ethical and legal principles into autonomous system design? The paper has argued that responsibility-aware design requires early incorporation of normative constraints, continuous adaptation to legal developments, and formal documentation of ethical trade-offs. While these questions have been addressed in depth, further nuance remains. For instance, the integration of XAI with

legal adjudication practices requires empirical validation, and the operationalization of fairness principles across cultural contexts needs further specification.

5.4 Limitations of the Current Research and Directions for Future Work

This research is bounded by several limitations. The scope was primarily conceptual and analytical which focused on high-level approaches rather than detailed empirical validation. While case studies provided grounding, they may not generalize across all domains or jurisdictions. The focus on formal methods and technical approaches may also underrepresent the sociopolitical dimensions of responsibility, including issues of power, institutional inertia, and public perception. Certain assumptions—such as the feasibility of integrating ethical constraints into autonomous systems—require more comprehensive validation through field deployment and longitudinal observation.

Future research should aim to bridge these gaps through interdisciplinary methodologies that combine technical modeling with insights from ethics, law, and social science. For instance:

a) Longitudinal studies could assess how responsibility attribution evolves in deployed systems over time.
b) Empirical work is needed to evaluate how different stakeholders interpret and respond to explainability features in high-stakes decisions.
c) Comparative legal analysis can illuminate how different jurisdictions conceptualize responsibility for non-human agents.
d) The development of responsibility reasoning benchmarks and simulation environments could support reproducible evaluation of proposed frameworks.

Thus, by expanding the empirical and interdisciplinary reach of the field, future work can build more resilient and context-aware models for reasoning about responsibility in autonomous systems.

REFERENCES

[1] V. Yazdanpanah *et al.*, "Reasoning about responsibility in autonomous systems: challenges and opportunities," *AI & Society*, vol. 38, no. 4, pp. 1453–1464, Dec. 2022, doi: 10.1007/s00146-022-01607-8.

[2] J. Sifakis and D. Harel, "Trustworthy autonomous system development," *ACM Transactions on Embedded Computing Systems*, vol. 22, no. 3, pp. 1–24, Jun. 2022, doi: 10.1145/3545178.

[3] M. Dastani and V. Yazdanpanah, "Responsibility of AI systems," *AI & Society*, vol. 38, no. 2, pp. 843–852, Jun. 2022, doi: 10.1007/s00146-022-01481-4.

[4] J. Dong, S. Chen, M. Miralinaghi, T. Chen, P. Li, and S. Labi, "Why did the AI make that decision? Towards an explainable artificial intelligence (XAI) for autonomous driving systems," *Transportation Research Part C Emerging Technologies*, vol. 156, p. 104358, Sep. 2023, doi: 10.1016/j.trc.2023.104358.

[5] K.-C. Hsu, H. Hu, and J. F. Fisac, "The Safety Filter: A Unified View of Safety-Critical Control in Autonomous Systems," *Annual Review of Control Robotics and Autonomous Systems*, vol. 7, no. 1, pp. 47–72, Feb. 2024, doi: 10.1146/annurev-control-071723-102940.

[6] M. Luckcuck, "Using formal methods for autonomous systems: Five recipes for formal verification," *Proceedings of the Institution of Mechanical Engineers Part O Journal of Risk and Reliability*, vol. 237, no. 2, pp. 278–292, Jul. 2021, doi: 10.1177/1748006x211034970.

[7] H.-M. Heyn, E. Knauss, and P. Pelliccione, "A compositional approach to creating architecture frameworks with an application to distributed AI systems," *Journal of Systems and Software*, vol. 198, p. 111604, Jan. 2023, doi: 10.1016/j.jss.2022.111604.

[8] N. H. Conradie, "Autonomous Military Systems: collective responsibility and distributed burdens," *Ethics and Information Technology*, vol. 25, no. 1, pp. 1–14, Feb. 2023, doi: 10.1007/s10676-023-09696-9.

[9] Jedličková, "Ethical approaches in designing autonomous and intelligent systems: a comprehensive survey towards responsible development," *AI & Society*, vol. 39, no. 4, pp. 1–14, Aug. 2024, doi: 10.1007/s00146-024-02040-9.

[10] M. M. Mayer, A. Buchner, and R. Bell, "Humans, machines, and double standards? The moral evaluation of the actions of autonomous vehicles, anthropomorphized autonomous vehicles, and human drivers in road-accident dilemmas," *Frontiers in Psychology*, vol. 13, pp. 1–13, Jan. 2023, doi: 10.3389/fpsyg.2022.1052729.

[11] M. A. Houghtaling *et al.*, "Standardizing an ontology for ethically aligned robotic and autonomous systems," *IEEE Transactions on Systems Man and Cybernetics Systems*, vol. 54, no. 3, pp. 1791–1804, Nov. 2023, doi: 10.1109/tsmc.2023.3330981.

[12] D. Trusilo, "Autonomous AI systems in Conflict: Emergent behavior and its impact on predictability and reliability," *Journal of Military Ethics*, vol. 22, no. 1, pp. 2–17, Jan. 2023, doi: 10.1080/15027570.2023.2213985.

[13] G. Bakirtzis, S. Carr, D. Danks, and U. Topcu, "Dynamic certification for autonomous systems," *Communications of the ACM*, vol. 66, no. 9, pp. 64–72, Aug. 2023, doi: 10.1145/3574133.

[14] F. Rowe, M. J. Medina, N. B. Journé, E. Coëtard, and M. Myers, "Understanding responsibility under uncertainty: A critical and scoping review of autonomous driving systems," *Journal of Information Technology*, vol. 39, no. 3, pp. 587–615, Sep. 2023, doi: 10.1177/02683962231207108.

[15] D. Saenz, Z. Harned, O. Banerjee, M. D. Abràmoff, and P. Rajpurkar, "Autonomous AI systems in the face of liability, regulations and costs," *Npj Digital Medicine*, vol. 6, no. 1, pp. 1–3, Oct. 2023, doi: 10.1038/s41746-023-00929-1.

[16] L. Methnani, M. Chiou, V. Dignum, and A. Theodorou, "Who's in charge here? a survey on trustworthy AI in variable autonomy robotic systems," *ACM Computing Surveys*, vol. 56, no. 7, pp. 1–32, Feb. 2024, doi: 10.1145/3645090.

[17] D. B. Acharya, K. Kuppan, and B. Divya, "Agentic AI: Autonomous Intelligence for Complex Goals – A Comprehensive survey," *IEEE Access*, vol. 13, pp. 1–25, Jan. 2025, doi: 10.1109/access.2025.3532853.

[18] H. Chakraa, F. Guérin, E. Leclercq, and D. Lefebvre, "Optimization techniques for Multi-Robot Task Allocation problems: Review on the state-of-the-art," *Robotics and Autonomous Systems*, vol. 168, p. 104492, Jul. 2023, doi: 10.1016/j.robot.2023.104492.

[19] X. Li, Z. Chen, J. M. Zhang, F. Sarro, Y. Zhang, and X. Liu, "Bias Behind the wheel: Fairness testing of autonomous driving systems," *ACM Transactions on Software Engineering and Methodology*, vol. 34, no. 3, Art. no. 82, Nov. 2024, doi: 10.1145/3702989.
[20] V. Jain, G. Gupta, M. Gupta, D. K. Sharma, and U. Ghosh, "Ambient intelligence-based multimodal human action recognition for autonomous systems," *ISA Transactions*, vol. 132, pp. 94–108, Nov. 2022, doi: 10.1016/j.isatra.2022.10.034.
[21] H. Araujo, M. R. Mousavi, and M. Varshosaz, "Testing, Validation, and Verification of Robotic and Autonomous Systems: A Systematic review," *ACM Transactions on Software Engineering and Methodology*, vol. 32, no. 2, pp. 1–61, Jun. 2022, doi: 10.1145/3542945.
[22] X. Yin, B. Gao, and X. Yu, "Formal synthesis of controllers for safety-critical autonomous systems: Developments and challenges," *Annual Reviews in Control*, vol. 57, p. 100940, Jan. 2024, doi: 10.1016/j.arcontrol.2024.100940.
[23] N. E. Fard, R. R. Selmic, and K. Khorasani, "Public Policy Challenges, regulations, oversight, technical, and ethical Considerations for autonomous Systems: a survey," *IEEE Technology and Society Magazine*, vol. 42, no. 1, pp. 45–53, Mar. 2023, doi: 10.1109/mts.2023.3241315.
[24] Holzinger, K. Zatloukal, and H. Müller, "Is Human Oversight to AI Systems still possible?," *New Biotechnology*, vol. 85, pp. 59–52, Dec. 2024, doi: 10.1016/j.nbt.2024.12.003.
[25] M. Adam, C. Diebel, M. Goutier, and A. Benlian, "Navigating autonomy and control in human-AI delegation: User responses to technology- versus user-invoked task allocation," *Decision Support Systems*, vol. 180, p. 114193, Feb. 2024, doi: 10.1016/j.dss.2024.114193.
[26] M. H. Cohen, T. G. Molnar, and A. D. Ames, "Safety-critical control for autonomous systems: Control barrier functions via reduced-order models," *Annual Reviews in Control*, vol. 57, p. 100947, Jan. 2024, doi: 10.1016/j.arcontrol.2024.100947.
[27] T. Ziemke, "Understanding Social robots: attribution of intentional agency to artificial and biological bodies," *Artificial Life*, vol. 29, no. 3, pp. 351–366, Jan. 2023, doi: 10.1162/artl_a_00404.
[28] K. Terzidis, F. Fabrocini, and H. Lee, "Unintentional intentionality: art and design in the age of artificial intelligence," *AI & Society*, vol. 38, no. 4, pp. 1715–1724, Jan. 2022, doi: 10.1007/s00146-021-01378-8.
[29] Rawal, A. Raglin, D. B. Rawat, B. M. Sadler, and J. McCoy, "Causality for Trustworthy Artificial intelligence: Status, challenges and perspectives," *ACM Computing Surveys*, vol. 57, no. 6, pp. 1–30, May 2024, doi: 10.1145/3665494.
[30] H. L. Fraser and N. P. Suzor, "Locating fault for AI harms: a systems theory of foreseeability, reasonable care and causal responsibility in the AI value chain," *Law Innovation and Technology*, vol. 17, no. 1, pp. 103–138, Mar. 2025, doi: 10.1080/17579961.2025.2469345.
[31] K. Shukla, V. Terziyan, and T. Tiihonen, "AI as a user of AI: Towards responsible autonomy," *Heliyon*, vol. 10, no. 11, p. e31397, May 2024, doi: 10.1016/j.heliyon.2024.e31397.
[32] P. Hacker, "The European AI liability directives – Critique of a half-hearted approach and lessons for the future," *Computer Law & Security Review*, vol. 51, p. 105871, Sep. 2023, doi: 10.1016/j.clsr.2023.105871.
[33] W. Xu, M. J. Dainoff, L. Ge, and Z. Gao, "Transitioning to Human Interaction with AI Systems: New Challenges and Opportunities for HCI Professionals to Enable Human-Centered AI," *International Journal of Human-Computer Interaction*, vol. 39, no. 3, pp. 494–518, Apr. 2022, doi: 10.1080/10447318.2022.2041900.
[34] V. Hassija *et al.*, "Interpreting Black-Box Models: A review on Explainable Artificial intelligence," *Cognitive Computation*, vol. 16, no. 1, pp. 45–74, Aug. 2023, doi: 10.1007/s12559-023-10179-8.
[35] S. Nazat, O. Arreche, and M. Abdallah, "On evaluating Black-Box explainable AI methods for enhancing anomaly detection in autonomous driving systems," *Sensors*, vol. 24, no. 11, p. 3515, May 2024, doi: 10.3390/s24113515.
[36] E. ŞAHiN, N. N. Arslan, and D. Özdemir, "Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning," *Neural Computing and Applications*, vol. 37, pp. 859–965, Nov. 2024, doi: 10.1007/s00521-024-10437-2.
[37] Kuznietsov, B. Gyevnar, C. Wang, S. Peters, and S. V. Albrecht, "Explainable AI for Safe and Trustworthy Autonomous Driving: A Systematic review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 12, pp. 19342–19364, Jan. 2024, doi: 10.1109/tits.2024.3474469.
[38] Novelli, M. Taddeo, and L. Floridi, "Accountability in artificial intelligence: what it is and how it works," *AI & Society*, vol. 39, no. 4, pp. 1871–1882, Feb. 2023, doi: 10.1007/s00146-023-01635-y.
[39] Y. Wang *et al.*, "Decision-Making Driven by driver intelligence and Environment Reasoning for High-Level Autonomous Vehicles: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 10, pp. 10362–10381, May 2023, doi: 10.1109/tits.2023.3275792.
[40] B. C. Stahl, "Embedding responsibility in intelligent systems: from AI ethics to responsible AI ecosystems," *Scientific Reports*, vol. 13, no. 1, pp. 1–8, May 2023, doi: 10.1038/s41598-023-34622-w.
[41] M. Taddeo, P. Jones, R. Abbas, K. Vogel, and K. Michael, "Socio-Technical Ecosystem Considerations: An Emergent Research Agenda for AI in Cybersecurity," *IEEE Transactions on Technology and Society*, vol. 4, no. 2, pp. 112–118, Jun. 2023, doi: 10.1109/tts.2023.3278908.
[42] Y. Himeur *et al.*, "AI-big data analytics for building automation and management systems: a survey, actual challenges and future perspectives," *Artificial Intelligence Review*, vol. 56, no. 6, pp. 4929–5021, Oct. 2022, doi: 10.1007/s10462-022-10286-2.
[43] T. Kim, M. D. Molina, M. Rheu, E. S. Zhan, and W. Peng, *One AI Does Not Fit All: A Cluster Analysis of the Laypeople's Perception of AI Roles*. ACM, 2023, pp. 1–20. doi: 10.1145/3544548.3581340.
[44] M. Hanna *et al.*, "Ethical and bias considerations in artificial intelligence (AI)/Machine learning," *Modern Pathology*, vol. 38, no. 3, p. 100686, Dec. 2024, doi: 10.1016/j.modpat.2024.100686.
[45] S. Albahri *et al.*, "A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion," *Information Fusion*, vol. 96, pp. 156–191, Mar. 2023, doi: 10.1016/j.inffus.2023.03.008.
[46] S. Grigorescu and M. Zaha, "CyberCortex.AI: An AI-based operating system for autonomous robotics and complex automation," *Journal of Field Robotics*, vol. 42, no. 2, pp. 474–492, Aug. 2024, doi: 10.1002/rob.22426.
[47] K. Grosse and A. Alahi, "A qualitative AI security risk assessment of autonomous vehicles," *Transportation Research Part C Emerging Technologies*, vol. 169, p. 104797, Sep. 2024, doi: 10.1016/j.trc.2024.104797.
[48] C. Sieber, L. M. V. Da Silva, K. Grünhagen, and A. Fay, "Rule-Based verification of autonomous unmanned aerial vehicles," *Drones*, vol. 8, no. 1, pp. 1–19, Jan. 2024, doi: 10.3390/drones8010026.
[49] P. H. Padovan, C. M. Martins, and C. Reed, "Black is the new orange: how to determine AI liability," *Artificial Intelligence and Law*, vol. 31, no. 1, pp. 133–167, Jan. 2022, doi: 10.1007/s10506-022-09308-9.
[50] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar, "A review of Trustworthy and Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 11, pp. 78994–79015, Jan. 2023, doi: 10.1109/access.2023.3294569.

[51] Taylor, "Is explainable AI responsible AI?," *AI & Society*, vol. 39, no. 2, pp. 1–10, Apr. 2024, doi: 10.1007/s00146-024-01939-7.
[52] M. Z. Zgurovsky, P. O. Kasyanov, and L. B. Levenchuk, "Formalization of methods for the development of autonomous artificial intelligence systems," *Cybernetics and Systems Analysis*, vol. 59, no. 5, pp. 763–771, Sep. 2023, doi: 10.1007/s10559-023-00612-z.
[53] S. Li, M. Wei, S. Li, and X. Yin, "Temporal logic task planning for autonomous systems with active acquisition of information," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 1436–1449, Oct. 2023, doi: 10.1109/tiv.2023.3327312.
[54] Z. Dimitrios and S. Petros, "Towards Responsible AI: A Framework for Ethical Design utilizing Deontic Logic," *International Journal of Artificial Intelligence Tools*, vol. 33, no. 8, p. 2550003, Jan. 2025, doi: 10.1142/s0218213025500034.
[55] Z. Guo and X. Liu, "How artificial intelligence cooperating with agent-based modeling for urban studies: A systematic review," *Transactions in GIS*, vol. 28, no. 3, pp. 654–674, Mar. 2024, doi: 10.1111/tgis.13152.
[56] C. B.-V. Burken, "Value Sensitive Design for autonomous weapon systems – a primer," *Ethics and Information Technology*, vol. 25, no. 1, pp. 1–14, Feb. 2023, doi: 10.1007/s10676-023-09687-w.
[57] Platzer, "Intersymbolic AI," in *Lecture notes in computer science*, vol. 15222, Springer LNCS, 2024, pp. 162–180. doi: 10.1007/978-3-031-75387-9_11.
[58] D. Garikapati and S. S. Shetiya, "Autonomous Vehicles: Evolution of artificial intelligence and the current industry landscape," *Big Data and Cognitive Computing*, vol. 8, no. 4, p. 42, Apr. 2024, doi: 10.3390/bdcc8040042.
[59] E. Papagiannidis, I. M. Enholm, C. Dremel, P. Mikalef, and J. Krogstie, "Toward AI governance: identifying best practices and potential barriers and outcomes," *Information Systems Frontiers*, vol. 25, no. 1, pp. 123–141, Apr. 2022, doi: 10.1007/s10796-022-10251-y.
[60] Q. Lu, L. Zhu, X. Xu, J. Whittle, D. Zowghi, and A. Jacquet, "Responsible AI Pattern Catalogue: A collection of best practices for AI governance and engineering," *ACM Computing Surveys*, vol. 56, no. 7, pp. 1–35, Oct. 2023, doi: 10.1145/3626234.
[61] Palaniappan, E. Y. T. Lin, and S. Vogel, "Global Regulatory Frameworks for the use of Artificial intelligence (AI) in the healthcare services sector," *Healthcare*, vol. 12, no. 5, p. 562, Feb. 2024, doi: 10.3390/healthcare12050562.
[62] N. Díaz-Rodríguez, J. Del Ser, M. Coeckelbergh, M. L. De Prado, E. Herrera-Viedma, and F. Herrera, "Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation," *Information Fusion*, vol. 99, p. 101896, Jun. 2023, doi: 10.1016/j.inffus.2023.101896.
[63] K. G. Hopster and M. M. Maas, "The technology triad: disruptive AI, regulatory gaps and value change," *AI And Ethics*, vol. 4, no. 4, pp. 1051–1069, Jun. 2023, doi: 10.1007/s43681-023-00305-5.
[64] I. Hauptman, B. G. Schelble, N. J. McNeese, and K. C. Madathil, "Adapt and overcome: Perceptions of adaptive autonomous agents for human-AI teaming," *Computers in Human Behavior*, vol. 138, p. 107451, Aug. 2022, doi: 10.1016/j.chb.2022.107451.
[65] N. Vallès-Peris and M. Domènech, "Caring in the in-between: a proposal to introduce responsible AI and robotics to healthcare," *AI & Society*, vol. 38, no. 4, pp. 1685–1695, Dec. 2021, doi: 10.1007/s00146-021-01330-w.
[66] C. Mennella, U. Maniscalco, G. De Pietro, and M. Esposito, "Ethical and regulatory challenges of AI technologies in healthcare: A narrative review," *Heliyon*, vol. 10, no. 4, p. e26297, Feb. 2024, doi: 10.1016/j.heliyon.2024.e26297.
[67] B. N. Thanh, H. X. Son, and D. T. H. Vo, "Blockchain: the economic and financial institution for autonomous AI?," *Journal of Risk and Financial Management*, vol. 17, no. 2, pp. 1–20, Jan. 2024, doi: 10.3390/jrfm17020054.
[68] Schmitt, "Automated machine learning: AI-driven decision making in business analytics," *Intelligent Systems With Applications*, vol. 18, p. 200188, Jan. 2023, doi: 10.1016/j.iswa.2023.200188.
[69] R. Hamon, H. Junklewitz, I. Sanchez, G. Malgieri, and P. De Hert, "Bridging the gap between AI and explainability in the GDPR: Towards Trustworthiness-by-Design in Automated Decision-Making," *IEEE Computational Intelligence Magazine*, vol. 17, no. 1, pp. 72–85, Jan. 2022, doi: 10.1109/mci.2021.3129960.
[70] W. Li, F. Paraschiv, and G. Sermpinis, "A data-driven explainable case-based reasoning approach for financial risk detection," *Quantitative Finance*, vol. 22, no. 12, pp. 2257–2274, Sep. 2022, doi: 10.1080/14697688.2022.2118071.
[71] S. He, "Who is Liable for the UBER Self-Driving Crash? Analysis of the Liability Allocation and the Regulatory Model for Autonomous Vehicles," in *Perspectives in law, business and innovation*, 1st ed., Springer, 2020, pp. 93–111. doi: 10.1007/978-981-15-9255-3_5.
[72] K. M. Bridges, "Race in the Machine: Racial Disparities in Health and Medical AI," *Virginia Law Review*, vol. 110, no. 2, pp. 1–97, Apr. 2024, [Online]. Available: https://virginialawreview.org/articles/race-in-the-machine-racial-disparities-in-health-and-medical-ai/
[73] B. L. Ranard *et al.*, "Minimizing bias when using artificial intelligence in critical care medicine," *Journal of Critical Care*, vol. 82, p. 154796, Mar. 2024, doi: 10.1016/j.jcrc.2024.154796.
[74] B. Vaassen, "AI, opacity, and personal autonomy," *Philosophy & Technology*, vol. 35, no. 4, pp. 1–20, Sep. 2022, doi: 10.1007/s13347-022-00577-5.
[75] Kumar, A. Aijaz, O. Chattar, J. Shukla, and R. Mutharaju, "Opacity, transparency, and the ethics of affective computing," *IEEE Transactions on Affective Computing*, vol. 15, no. 1, pp. 4–17, May 2023, doi: 10.1109/taffc.2023.3278230.

BIOGRAPHIES OF AUTHORS

Usman Tariq is an Associate Professor at Prince Sattam Bin Abdulaziz University, Saudi Arabia. He earned his Ph.D. from Ajou University, South Korea, in 2010. With a research portfolio exceeding $1 million in funding, he has made impactful contributions in the fields of wireless sensor networks, IoT systems, cybersecurity, and intelligent infrastructure. His work has led to the publication of over 200 research articles. Widely regarded as an authority in his domain, Dr. Tariq has chaired international conferences and delivered more than 50 keynote addresses and invited talks. He can be reached at u.tariq@psau.edu.sa.

Irfan Ahmed is the Engineering Foundation Endowed Associate Professor of Computer Science at Virginia Commonwealth University (VCU), where he also serves as a NIRA Scholar, a Fellow of the American Academy of Forensic Sciences, and a Faculty Fellow at the VCU Cybersecurity Center. He leads the Security and Forensics Engineering (SAFE) Research Lab, focusing on digital forensics, malware analysis, and cyber-physical system security. A recipient of numerous national awards—including honors from ORAU, USCYBERCOM, and AAFS—his work has been funded by agencies such as NSF, NSA, DOE, and ONR, and has earned multiple Best Paper and Poster Awards at top cybersecurity conferences. He can be reached at iahmed3@vcu.edu