

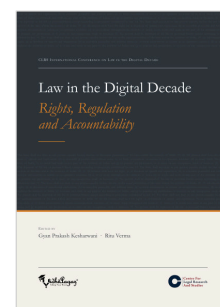
CHAPTER 10

Beyond the Black Box: A Multi Tiered Framework for Algorithmic Accountability and Liability in Autonomous Decision Making Systems

Lakshmi Niveditha S and Anna Mariam Iype

in *Law in the Digital Decade: Rights, Regulation and Accountability*, edited by Gyan Prakash Kesharwani and Ritu Verma (VidhiAagaz 2026), pp. 111–123

DOI <https://doi.org/10.63108/VAB.LDD.1.10>



CITATIONS

BLUEBOOK (21ST ED.)

Lakshmi Niveditha S & Anna Mariam Iype, *Beyond the Black Box: A Multi Tiered Framework for Algorithmic Accountability and Liability in Autonomous Decision Making Systems*, in *LAW IN THE DIGITAL DECADE: RIGHTS, REGULATION AND ACCOUNTABILITY* 111 (Gyan Prakash Kesharwani & Ritu Verma eds., VidhiAagaz 2026).

OSCOLA (4TH ED.)

Lakshmi Niveditha S and Anna Mariam Iype, 'Beyond the Black Box: A Multi Tiered Framework for Algorithmic Accountability and Liability in Autonomous Decision Making Systems' in Gyan Prakash Kesharwani and Ritu Verma (eds), *Law in the Digital Decade: Rights, Regulation and Accountability* (VidhiAagaz 2026) 111.

CHICAGO (17TH ED.)

Niveditha, Lakshmi S, and Anna Mariam Iype. "Beyond the Black Box: A Multi Tiered Framework for Algorithmic Accountability and Liability in Autonomous Decision Making Systems." In *Law in the Digital Decade: Rights, Regulation and Accountability*, edited by Gyan Prakash Kesharwani and Ritu Verma, 111–123. Maihar: VidhiAagaz, 2026. <https://doi.org/10.63108/VAB.LDD.1.10>.

APA (7TH ED.)

Niveditha, L. S., & Iype, A. M. (2026). Beyond the black box: A multi tiered framework for algorithmic accountability and liability in autonomous decision making systems. In G. P. Kesharwani & R. Verma (Eds.), *Law in the digital decade: Rights, regulation and accountability* (pp. 111–123). VidhiAagaz. <https://doi.org/10.63108/VAB.LDD.1.10>

SOURCE

<https://doi.org/10.63108/VAB.LDD.1.10>

LICENCE

Open access under the Creative Commons Attribution-NonCommercial 4.0 International Licence (CC BY-NC 4.0): it may be copied, shared and adapted for non-commercial purposes, with credit to the authors and the source.

Citations are given as a guide; check them against the style manual you follow. This page was added by VidhiAagaz to the downloaded file. The chapter follows, with the page numbers of the printed book.

CHAPTER 10

Beyond the Black Box: A Multi Tiered Framework for Algorithmic Accountability and Liability in Autonomous Decision Making Systems

LAKSHMI NIVEDITHA S* AND ANNA MARIAM IYPE**

As artificial intelligence systems become increasingly capable of making decisions with limited human intervention, questions of responsibility and accountability have become more difficult. This paper examines two major challenges created by autonomous decision-making systems: the “black box” problem, where the reasoning behind an AI decision may be difficult to understand, and the “many hands” problem, where responsibility is spread across several actors involved in the development and use of the system. The paper further uses actor mapping to identify the different participants involved in the AI lifecycle and examines why identifying these actors alone may not be enough to determine responsibility.

To address this gap, the paper proposes a Multi-Tiered Accountability Framework based on three connected levels: technical accountability, actor-based accountability and responsibility allocation. The framework considers factors such as an actor’s contribution, level of control, foreseeability of risk and ability to prevent or reduce harm. It aims to provide a clearer and fairer approach to accountability without treating AI autonomy as a reason to remove human responsibility. The paper ultimately argues that as AI becomes more autonomous, legal accountability must develop alongside technological innovation.

Keywords: Artificial Intelligence; Autonomous Decision-Making Systems; Algorithmic Accountability; Black Box Problem; Many Hands Problem; Actor Mapping; AI Liability; Human Oversight

1 INTRODUCTION

Artificial intelligence has moved beyond being a tool that simply follows instructions given by humans. AI systems are increasingly capable of analysing large amounts of information, identifying patterns, making predictions and producing decisions with limited human intervention. The European Union’s AI Act itself recognises that AI systems may operate with different levels of autonomy and may generate outputs such as predictions, recommendations and decisions that influence physical or virtual environments.¹ This increasing autonomy creates an important legal question: When an AI system causes harm, who should be held responsible?

The question becomes more difficult because an AI system does not operate in isolation. Behind a single AI-generated decision may be a developer who designed the system, a company that supplied or trained it, a data provider, an organisation that deployed it, and a human operator who relied upon its output. Therefore, simply asking who “made” the decision may not always provide a satisfactory answer.

* Student at Christ (Deemed to be University), Bengaluru, Karnataka, India.

** Student at Christ (Deemed to be University), Bengaluru, Karnataka, India.

¹ Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence, art 3(1).

The problem becomes even more serious when the reasoning behind an AI decision cannot be clearly understood. Machine-learning systems may process enormous quantities of data and identify patterns that are difficult for even their creators or users to follow. This is commonly referred to as the “black box” problem. The concern is not merely that AI is complicated, but that a person affected by an AI decision may be unable to understand why that decision was reached or who should explain it.²

At the same time, responsibility may be spread across several human and institutional actors. This creates what has been described as the “many hands” problem, where several actors contribute to an outcome but no single actor appears to have complete responsibility for it.³ In the context of AI, the developer may argue that the system was used differently from its intended purpose, while the user may argue that the system was designed by someone else. The result can be a gap between the occurrence of harm and the identification of a responsible party.

This paper therefore proceeds from the position that AI should not simply be restricted because it is becoming more autonomous. Technological development is inevitable, and the law cannot realistically expect technology to remain within the boundaries of older legal categories. However, allowing technology to develop does not mean allowing accountability to disappear with it.

The central argument of this paper is that responsibility for autonomous decision-making systems should be determined through a multi-tiered framework. Rather than automatically placing responsibility upon the developer, manufacturer, user or organisation, the law should examine the different actors involved, their contribution, their degree of control, the risks they were capable of identifying, and the safeguards they were expected to maintain. This approach begins with understanding the nature of autonomous decision-making systems, followed by an examination of the black box and many hands problems, before mapping the actors involved in the AI lifecycle.

2 AUTONOMOUS DECISION-MAKING SYSTEMS

The term “autonomous decision-making system” refers to an AI-based system that is capable of processing information and producing outputs without requiring a human being to make every individual decision. Autonomy, however, should not be understood as meaning that the AI system has become a completely independent human-like actor. Rather, it refers to the degree to which the system can operate without direct human intervention at each stage.

This distinction is important because not every automated system creates the same legal problem. A simple system that follows a fixed instruction is relatively easier to understand. For example, a calculator will produce an answer based on a mathematical formula. A more advanced machine-learning system, however, may identify patterns from data and produce an outcome that was not individually programmed by a human. The more such systems are capable of adapting, learning or making decisions with limited human intervention, the more difficult it becomes to trace the exact chain that led to a particular outcome.

The growing use of large datasets is one of the reasons for this difficulty. Modern AI systems can process extremely large amounts of information and identify patterns across that information. This ability is one of their greatest advantages, but it can also create risks. If the information used to train or operate

² Jenna Burrell, ‘How the Machine “Thinks”: Understanding Opacity in Machine Learning Algorithms’ (2016) 3 *Big Data & Society* 1.

³ Filippo Santoni de Sio and Giulio Mecacci, ‘Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them’ (2021) 35 *Philosophy & Technology* 1057.

an AI system contains bias, inaccuracies or incomplete information, the resulting system may reproduce or even amplify those problems.

Therefore, the question is not simply whether AI is “intelligent”. The more important question from a legal perspective is “What happens when an AI system acts on the basis of its own processing and that action causes harm?”

For instance, consider an AI system used by an organization to assess applications for employment, loans, insurance or other services. If the system produces an unfavorable decision, the affected individual may naturally ask why the decision was made. If a human employee makes the decision, it is generally possible to identify the person, ask for the reasoning behind the decision and examine whether that person acted negligently, unfairly or unlawfully. With an autonomous AI system, however, the process may involve several stages of data collection, training, programming, deployment and operation.

This creates what can broadly be described as a responsibility gap. Scholars have argued that increasing autonomy can make it difficult to determine whether responsibility lies with the programmer, manufacturer, operator or another actor when an autonomous system produces an undesirable outcome.⁴ At the same time, it would be too simplistic to assume that every autonomous system automatically creates a responsibility gap. Whether such a gap exists depends upon the particular circumstances, including the nature of the system, the degree of human control and the roles played by different actors.⁵

This is where the present research takes a middle position. AI systems should not be treated as inherently unsafe merely because they are autonomous. At the same time, autonomy should not become an excuse for avoiding human responsibility. The law must instead examine the circumstances surrounding each system and determine which actors were in a position to prevent, control or respond to the harm.

Autonomy, therefore, creates the starting point of the accountability problem, but it does not by itself answer the question of liability. To answer that question, it is necessary to examine what happens inside the system and how its decision-making process can be understood.

2.1 The Black Box Problem

One of the most important challenges created by AI-based decision-making is the black box problem. In simple terms, a black box refers to a system where the relationship between the information given to the system and the final output is difficult to understand or explain.

The problem is particularly significant in machine-learning systems. Unlike traditional computer programs, where a programmer may directly specify the rules that determine the output, machine-learning systems can learn patterns from data. As a result, even when the system is functioning according to its technical design, it may be difficult for a human being to explain exactly why a particular result was produced.

Jenna Burrell identifies different forms of algorithmic opacity, including secrecy, lack of technical understanding and opacity arising from the complexity and scale of machine-learning systems themselves.⁶ This distinction is important because the black box problem does not always mean that a company

⁴ Peter Königs, ‘Artificial Intelligence and Responsibility Gaps: What Is the Problem?’ (2022) 24 *Ethics and Information Technology* 28.

⁵ Ibid.

⁶ Burrell (n 2).

is deliberately hiding information. Sometimes, the problem is that the process is technically difficult to interpret even for people involved in developing or operating the system.

This creates a serious problem when AI is used in areas affecting people's rights and interests. If an AI system rejects a loan application, identifies a person as a security risk, rejects a job application or makes another consequential classification, the affected individual may need to know the basis of the decision. Without meaningful information about how the decision was reached, challenging the decision becomes considerably more difficult.

The problem becomes even more significant when responsibility is being considered. Suppose an AI system produces a harmful decision and the developers themselves cannot fully explain why the system produced that result. The question then becomes: How can responsibility be fairly assigned?

This does not necessarily mean that such systems should never be used. Instead, they should be used with an appropriate level of caution, particularly where the consequences of error are serious. The National Institute of Standards and Technology (NIST), for example, treats accountability and transparency as important characteristics of trustworthy AI and links meaningful transparency to information concerning design decisions, training data, model training, system structure, intended uses and decisions made during deployment.⁷

This demonstrates why transparency cannot simply mean revealing the source code. A person affected by an AI decision may not benefit from being given thousands of lines of code if they still cannot understand how that code affected the outcome. Meaningful accountability requires information that is relevant to the particular stage of the AI lifecycle and the people interacting with the system.

The black box problem also creates an evidentiary difficulty. In a traditional dispute, the parties may attempt to establish what happened by examining documents, decisions and the conduct of the relevant person. With an AI system, the relevant evidence may be spread across training data, model design, system logs, deployment decisions and human interactions with the system.

This is why explainability and transparency are important, but they should not be treated as complete solutions to the problem. Making an algorithm more explainable may help an affected person understand the decision, but it does not automatically tell us who should be legally responsible for the consequences. Responsibility may still be distributed among several actors.

The black box therefore represents one side of the accountability problem: we may not understand how the decision was produced. The many hands problem represents the other side: even if we understand how the system was created and used, several different actors may have contributed to the outcome.

2.2 The Many Hands Problem

The second major difficulty is the many hands problem. Unlike the black box problem, which focuses mainly on the opacity of the decision-making process, the many hands problem focuses on the distribution of responsibility between different actors.

An AI system is rarely created, controlled and used by one person. A single system may involve software developers, data scientists, data providers, manufacturers, technology companies, organisations that purchase or deploy the system, human operators and end users. Regulators may also influence how the system can be developed and used.

⁷ National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1, 2023); see also NIST, 'AI Risks and Trustworthiness'.

Recent scholarship specifically connects AI responsibility gaps with this problem of responsibility distribution. Where several agents contribute to an outcome, identifying the relevant actors is only the first step; determining how much responsibility each actor should bear remains a separate question.⁸

Consider a hypothetical AI system used to decide whether individuals qualify for a particular service. One company develops the underlying model. Another supplies the data. A third organisation integrates the system into its own platform. A fourth organisation deploys it in practice, while employees use the system's output to make decisions affecting individuals.

If the system produces a discriminatory or otherwise harmful result, it would be unreasonable to automatically conclude that the final user alone is responsible. At the same time, it may also be unfair to automatically blame the original developer without examining how the system was deployed or whether the user ignored warnings and safeguards.

Therefore, responsibility should be divided according to the contribution, control and circumstances of each actor.

This does not mean that every person involved should automatically share equal liability. Equal participation does not necessarily mean equal responsibility. A developer who knowingly creates a defective system, a company that knowingly deploys it in a high-risk environment and an employee who unknowingly relies on its output may have very different levels of responsibility.

The appropriate approach must therefore remain contextual. The same AI system could produce different liability questions depending on how it was used, what information was available to the relevant actors, what safeguards existed and what degree of control each actor possessed.

At the same time, the rules used to make these assessments should not change unpredictably from one situation to another. There needs to be a stable framework through which courts and other decision-makers can assess the roles of different actors. Otherwise, uncertainty itself may become a barrier to responsible AI development.

This is where the many hands problem becomes particularly important to the present research. It demonstrates why simply identifying “a human behind the AI” is not sufficient. There may be many humans behind the AI, each performing a different function.

The challenge is therefore to move from the broad question “Who is responsible?” to the more useful questions:

1. Who created or designed the system?
2. Who supplied or controlled the relevant data?
3. Who decided to deploy the system?
4. Who had control over its use?
5. Who could reasonably have identified the risk?
6. Who had the ability to prevent or reduce the harm?
7. Who failed to take an appropriate safeguard?

Answering these questions provides the basis for a more structured approach to accountability.

⁸ Filippo Santoni de Sio and Giulio Mecacci, ‘Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them’, *Philosophy & Technology* 34, no.4.

2.2.1 Actor Mapping

Before liability can be fairly assigned, the actors involved in an autonomous decision-making system must first be identified. This process can be described as actor mapping.

Actor mapping does not mean deciding in advance who is legally liable. Instead, it creates a structured picture of the different participants involved in the AI lifecycle. This is important because responsibility may arise at different stages and for different reasons.

TABLE 1. *Actors in the AI lifecycle and possible accountability concerns*

ACTOR	ROLE IN AI LIFECYCLE	POSSIBLE ACCOUNTABILITY CONCERN
Developer	Designs and develops the AI system	Design defects, inadequate safeguards or foreseeable risks
Data provider	Supplies training or operational data	Poor-quality, inaccurate or biased data
Manufacturer	Integrates the AI system into a product or service	Integration or technical failures
Deployer	Introduces the system into a real-world environment	Inappropriate deployment or inadequate monitoring
Operator/User	Uses or supervises the system	Misuse, failure to follow safeguards or unreasonable reliance
Organization	Determines how and why the system is used	Governance, oversight and risk-management failures
Regulator	Establishes and enforces applicable standards	Regulatory oversight and compliance

The purpose of this map is not to suggest that each actor will necessarily be liable whenever something goes wrong. Instead, it helps identify where responsibility could potentially arise.

For example, if an AI system produces a harmful result because its training data contained a serious and foreseeable bias, the data-related actor and developer may require examination. If the system itself was adequately designed but an organisation deployed it in a context for which it was never intended, the focus may shift towards the deployer. Similarly, if adequate safeguards existed but an operator deliberately ignored them, the operator's conduct may become relevant.

This approach also reflects the fact that accountability exists throughout the AI lifecycle rather than only at the moment when the final decision is made. NIST's AI Risk Management Framework similarly emphasises the importance of identifying roles and responsibilities across the AI lifecycle and treating AI risk management as an ongoing process rather than a one-time exercise.⁹

Actor mapping therefore acts as a bridge between the conceptual problems identified above and the question of legal liability. The black box problem tells us that we may struggle to understand how the system reached a particular outcome. The many hands problem tells us that several actors may have contributed to that outcome. Actor mapping provides a method of identifying those actors and examining their respective roles.

The central principle emerging from this analysis is that accountability should follow meaningful involvement and control rather than simply follow ownership of the AI system. A person should not escape responsibility merely because the final decision was technically produced by an algorithm. At the same time, an individual should not automatically be blamed merely because they were the last person to interact with the system.

⁹ National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1, 2023).

Ultimately, the law must distinguish between different forms and degrees of involvement. The greater an actor's ability to anticipate, control, prevent or mitigate a particular risk, the stronger the basis for examining that actor's responsibility. This does not create an automatic rule of liability, but it provides a rational starting point for determining where liability should be placed.

This actor-based approach forms the foundation for the subsequent stages of the proposed multi-tiered framework, which examine how responsibility can be translated into appropriate forms of accountability and liability.

3 ACCOUNTABILITY THEORY

Accountability is a fundamental concept in understanding how power and decision-making should be exercised, particularly where decisions can affect individuals and society. Traditionally, accountability refers to a relationship in which an actor is required to explain and justify their conduct to a forum that has the authority to question, evaluate and potentially impose consequences.¹⁰ This traditional understanding assumes that the person or institution making a decision can be identified and held answerable for that decision.

With the increasing use of artificial intelligence and automated decision-making systems, this traditional understanding of accountability has become more complicated. Algorithmic accountability extends the concept of accountability to systems in which decisions may be influenced or produced by algorithms rather than solely by human decision-makers. It therefore requires attention not only to the final outcome but also to the design, development, deployment and use of the AI system.¹¹

Accountability should also be distinguished from related concepts such as transparency and explainability. Transparency generally concerns access to information about how a system operates, while explainability focuses on making the reasoning or output of an AI system understandable. Accountability is broader because it concerns who must answer for a decision, to whom they must answer, according to what standards, and what consequences may follow. Thus, transparency or explainability may support accountability, but they do not by themselves establish accountability.

Similarly, accountability, responsibility and legal liability are not interchangeable concepts. Responsibility concerns who is expected to perform or control an action, while accountability involves being answerable for conduct and its consequences. Legal liability, on the other hand, concerns whether a person or organisation can be held legally responsible under a particular legal rule. Keeping these concepts separate is particularly important in autonomous AI systems because identifying the legally liable party may become difficult when decisions result from complex interactions between developers, deployers, users and the system itself.

This difficulty is reflected in the idea of a responsibility gap. Andreas Matthias argues that learning and autonomous systems can behave in ways that could not have been fully predicted by their designers or operators. As a result, traditional approaches to assigning responsibility may struggle to identify a clear

¹⁰ Mark Bovens, 'Analysing and Assessing Accountability: A Conceptual Framework' (2007) 13(4) *European Law Journal* 447.

¹¹ Claudio Novelli, Mariarosaria Taddeo and Luciano Floridi, 'Accountability in Artificial Intelligence: What It Is and How It Works' (2024) 39 *AI & Society* 1871.

human actor responsible for a particular outcome.¹² In AI decision-making, responsibility may therefore be distributed across several actors rather than resting with a single individual.

The problem becomes more significant as AI systems gain greater autonomy. Where human intervention is limited, it becomes necessary to examine whether meaningful human control existed over the system and its decisions. The concept of meaningful human control therefore provides an important basis for determining how human actors should remain connected to autonomous systems and their outcomes.¹³

These difficulties can ultimately produce what may be described as an accountability gap—a situation in which an AI-generated decision has significant consequences, but no actor can be clearly identified as answerable for the decision. The complexity of algorithmic systems, combined with their capacity to operate with limited human intervention, can make conventional accountability mechanisms inadequate.

A human-centred approach to accountability seeks to address this problem by ensuring that accountability remains attached to human and institutional actors rather than treating the AI system itself as the accountable party. This requires identifying the relevant actors throughout the AI lifecycle and establishing mechanisms through which their decisions and actions can be questioned and evaluated.

A further approach is to understand accountability as distributed across the AI lifecycle. Instead of assigning accountability only at the point where an automated decision produces harm, responsibility and answerability can be considered at different stages, including system design, development, deployment, monitoring and use. This approach recognises that different actors may exercise different forms of control over an AI system and may therefore have different accountability obligations.

Overall, the accountability framework for autonomous decision-making systems must move beyond the idea of identifying a single decision-maker. Traditional accountability theory provides the foundation, while algorithmic accountability adapts that foundation to systems involving multiple actors, limited predictability and varying degrees of human control. The responsibility gap and the need for meaningful human control demonstrate why accountability must be distributed across the AI lifecycle rather than concentrated solely on the final user or decision-maker.

This theoretical understanding of accountability provides the foundation for examining the next issue: legal liability. While accountability determines who should be answerable for AI-related decisions, legal liability asks a more specific question—who can actually be held legally responsible when an autonomous AI system causes harm or violates a legal right?

4 LEGAL LIABILITY IN AUTONOMOUS AI DECISION-MAKING

Legal liability refers to the legal consequences that arise when an individual or organization breaches a legal duty and causes legally recognized harm. In the context of artificial intelligence, the question of liability becomes more complex because the harmful outcome may result not from the direct action of a single person, but from the interaction between developers, providers, deployers, users and an increasingly autonomous system. It is therefore important to distinguish liability from accountability and responsibility. While accountability concerns the obligation to explain and justify conduct, and responsibility

¹² Andreas Matthias, 'The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata' (2004) 6 *Ethics and Information Technology* 175.

¹³ Filippo Santoni de Sio and Jeroen van den Hoven, 'Meaningful Human Control over Autonomous Systems: A Philosophical Account' (2018) 5 *Frontiers in Robotics and AI* 15.

may refer to the duties or obligations associated with an actor's role, legal liability determines whether an actor can be held legally responsible and required to provide a remedy for harm. The distinction becomes particularly significant where an AI system produces an outcome that no individual actor directly intended or fully anticipated.¹⁴

4.1 Attribution of Liability

The central problem in AI liability is therefore one of attribution: identifying the person or legal entity to whom a harmful AI decision can properly be attributed. An AI system does not operate in isolation; its behavior may reflect decisions made during its design, training, deployment, monitoring and use. Consequently, liability may potentially involve the developer who designed the system, the provider who placed it on the market, the organization that deployed it, or the operator who relied upon its output. The difficulty lies in determining which actor exercised sufficient control over the relevant risk to justify imposing legal liability. This is particularly challenging where several actors contribute to the operation of the system and their respective roles overlap.¹⁵

4.2 The Liability Gap in Autonomous AI

Autonomous and adaptive AI systems create what has been described as a liability gap. Traditional liability rules generally operate on the assumption that a harmful act can be connected to an identifiable human actor, a breach of duty or a defective product. However, autonomous systems may produce outcomes that are difficult to predict even by their designers or operators. This does not mean that the AI system itself should automatically become the bearer of legal liability; rather, it exposes a gap between the occurrence of harm and the ability of existing legal rules to identify an appropriate defendant. The more autonomous, complex and unpredictable the system becomes, the more difficult it may be to determine where one actor's legal responsibility ends and another's begins.¹⁶

4.3 Fault-Based Liability and Negligence

One possible route for addressing AI-related harm is fault-based liability, particularly through negligence. Traditionally, negligence requires the existence of a duty of care, a breach of that duty, a causal connection between the breach and the harm, and legally recognized damage. These principles can potentially apply to AI systems where, for example, a developer fails to conduct adequate testing or a deployer fails to exercise reasonable supervision. However, their application becomes difficult where the specific harmful behaviour of an autonomous system could not reasonably have been anticipated. The concept of foreseeability, which is central to many negligence-based approaches, becomes particularly difficult to apply when an AI system can learn from new data, adapt after deployment or generate unexpected outputs.¹⁷

¹⁴ Mark Bovens, 'Analysing and Assessing Accountability: A Conceptual Framework' (2007) 13(4) *European Law Journal* 447.

¹⁵ Miriam C Buiten, 'Product Liability for Defective AI' (2024) 57 *European Journal of Law and Economics* 239.

¹⁶ Simon Burton and others, 'Mind the Gaps: Assuring the Safety of Autonomous Systems from an Engineering, Ethical, and Legal Perspective' (2020) 279 *Artificial Intelligence* 103201; see also Miriam C Buiten, Alexandre de Streel and Martin Peitz, 'The Law and Economics of AI Liability' (2023) 48 *Computer Law & Security Review* 105794.

¹⁷ *Ibid.*

4.4 Strict Liability and Product Liability

These difficulties raise the question of whether strict or no-fault forms of liability may be more appropriate for certain high-risk AI applications. Under a strict-liability approach, the claimant does not necessarily have to establish the same level of fault or negligence on the part of the defendant. Such an approach could provide stronger protection where the risks associated with highly autonomous systems are significant but difficult to predict. Product liability is particularly relevant because AI may form part of a physical product or operate as software in its own right. Modern product-liability approaches increasingly recognise that software and AI can create legally relevant risks. For example, the EU's revised Product Liability Directive expressly treats software, including AI systems, as a product for the purposes of its no-fault product-liability regime.¹⁸

However, product liability does not completely resolve the problem. AI systems may change through updates, learning processes or interaction with their environment, making it difficult to determine whether the relevant defect existed when the product entered the market. The question therefore becomes not simply whether an AI system was defective, but who had control over the relevant risk and at what stage of the system's lifecycle that risk arose. This makes the allocation of liability between producers, providers and users particularly important.¹⁹

4.5 Allocation of Liability Across the AI Lifecycle

Rather than treating AI liability as a question of identifying one person to blame, a more appropriate approach may be to consider how liability should be allocated across the AI lifecycle. Developers may be responsible for design and development failures; providers may bear obligations concerning the system as supplied; deployers may be responsible for inappropriate implementation or inadequate monitoring; and users or operators may be liable where they misuse the system or disregard required safeguards. This approach recognises that control over an AI system is often distributed rather than concentrated in one actor. It also avoids automatically transferring liability to the person who happens to be closest to the final harmful decision when the underlying risk may have originated much earlier in the system's development.²⁰

4.6 Human Oversight and Remedies

Human oversight therefore remains an important consideration when determining liability for autonomous decision-making. Where a system is capable of producing significant consequences, the presence or absence of meaningful human supervision may influence whether a deployer or operator has fulfilled its legal duties. Effective oversight should not merely consist of having a human formally present; it should involve the ability to understand relevant limitations, monitor the system and intervene when necessary. The EU AI Act, for example, requires appropriate human oversight for high-risk AI systems

¹⁸ Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products. The Directive expressly includes software, including AI systems, within its product-liability framework.

¹⁹ Ibid.

²⁰ Bart Custers, Henning Lahmann and Benjamyn I Scott, 'From Liability Gaps to Liability Overlaps: Shared Responsibilities and Fiduciary Duties in AI and Other Complex Technologies' (2025) 40 *AI & Society* 4035.

and recognises the need for humans to monitor, interpret and, where appropriate, override or stop the system.²¹

Ultimately, the objective of an AI liability framework should not simply be to identify a defendant after harm has occurred. It should also ensure that individuals affected by autonomous decision-making have a meaningful avenue for redress and compensation, while creating incentives for developers, providers and deployers to manage foreseeable risks before harm occurs. The central challenge is therefore to construct a liability model that reflects the distributed nature of AI development and use without allowing autonomy or technological complexity to become a means of escaping legal responsibility.

5 RESEARCH GAP

Existing discussions on AI accountability have largely focused on individual concerns such as transparency, explainability, human oversight, bias and liability. These issues are important, but they are often considered separately. The black box problem explains why it may be difficult to understand how an AI system reaches a decision, while the many hands problem explains why it may be difficult to identify who should be held responsible for that decision.

The difficulty becomes greater when these problems occur together. Even if an AI system can be examined technically, this does not automatically tell us which actor should be legally responsible when harm occurs. Similarly, identifying all the people and organisations involved in an AI system does not mean that all of them should be treated as equally responsible.

There is therefore a gap between understanding how an AI system operates and determining who should be accountable for its outcomes. A clear framework is needed to connect these two aspects. Such a framework should consider not only the technical functioning of the system, but also the role, contribution, control and ability to prevent harm of the different actors involved.

This paper addresses this gap by proposing a Multi-Tiered Accountability Framework. The framework aims to ensure that increasing autonomy in AI systems does not result in a corresponding loss of accountability.

6 MULTI-TIERED ACCOUNTABILITY FRAMEWORK

The proposed framework consists of three connected tiers: technical accountability, actor accountability and responsibility allocation. The tiers are designed to work together rather than as separate stages. The first helps understand the system, the second identifies the actors involved, and the third determines how responsibility should be allocated.

6.1 Tier One: Technical Accountability

The first tier focuses on the AI system itself. Before responsibility can be meaningfully discussed, there must be enough information to understand what happened within the system and how it was developed and deployed.

This does not mean that every AI system must be completely explainable in simple human terms. Complex machine-learning systems may not always allow their decisions to be reduced to a simple chain

²¹ Regulation (EU) 2024/1689 (Artificial Intelligence Act), art 14.

of reasoning. However, there should be sufficient documentation and monitoring to examine important aspects such as the data used, the purpose of the system, its design, known limitations, testing processes and significant changes made after deployment.

This tier directly addresses the black box problem. When an AI system produces a harmful result, complete opacity can make it difficult to determine whether the problem arose from the design of the system, the data used to train it, the way it was deployed or the way it was operated.

Therefore, technical accountability should involve appropriate testing, documentation, monitoring and record keeping. The question should not only be “Can we explain exactly why the AI made this decision?”, but also “Do we have enough information to investigate what happened?” This makes technical transparency an important foundation for accountability.

6.2 Tier Two: Actor Accountability

The second tier addresses the many hands problem. Autonomous AI systems are usually not created, controlled or used by one person alone. Developers, data providers, manufacturers, technology companies, organisations, operators and other actors may all have different roles in the system.

The purpose of this tier is therefore to identify these actors and understand their individual roles. Actor mapping should consider who designed the system, who supplied or controlled the data, who decided to deploy it, who operated it and who was responsible for monitoring it.

However, identifying an actor does not automatically make that actor liable. The framework recognises that different actors may have different levels of responsibility.

For example, a developer may be responsible for a defect in the design, while an organisation that deploys the system may be responsible for using it in an unsuitable context. An operator may also bear responsibility where they had the ability to intervene but failed to do so.

The framework therefore asks:

1. Who created or developed the system?
2. Who supplied or controlled the relevant data?
3. Who decided to deploy the system?
4. Who had control over its use?
5. Who could reasonably have identified the relevant risk?
6. Who had the ability to prevent or reduce the harm?

These questions help ensure that responsibility is not placed automatically on the person who happened to be closest to the final decision.

6.3 Tier Three: Responsibility Allocation

The final tier determines how responsibility should actually be distributed between the actors identified in the previous tier.

The framework proposes that responsibility should be based on four main considerations: contribution, control, foreseeability and prevention.

First, contribution considers what role an actor played in creating or contributing to the harmful outcome. Second, control considers the extent to which the actor could influence the relevant part of the

AI system. Third, foreseeability considers whether the risk could reasonably have been identified. Finally, prevention considers whether the actor had a realistic opportunity to prevent or reduce the harm.

These factors allow the framework to recognise shared but differentiated responsibility. More than one actor may be responsible for an outcome, but this does not mean that every actor should bear the same degree of responsibility.

For example, if a developer releases a system with a serious and foreseeable defect, the developer may bear greater responsibility. If an organization knowingly deploys the system for a purpose for which it was not designed, responsibility may instead shift towards the deployer. Similarly, where meaningful human oversight existed but was ignored, the failure of that oversight may also become relevant.

The framework therefore avoids two extremes: treating AI as completely independent and placing responsibility on no one, or treating one human actor as automatically responsible for everything the AI does.

The underlying principle is simple: the more control and ability an actor had to identify, prevent or reduce a particular risk, the stronger the basis for holding that actor responsible for the resulting harm.

7 CONCLUSION

Autonomous decision-making systems are becoming increasingly capable of performing tasks that were once carried out directly by humans. While this development can provide significant benefits, it also creates new challenges for accountability. The difficulty is not simply that AI can make decisions autonomously, but that those decisions may be difficult to understand and may involve several different actors.

The black box problem creates uncertainty about how an AI system reaches its decisions, while the many hands problem creates uncertainty about who should be held responsible when those decisions cause harm. Actor mapping helps identify the participants in the system, but identifying them alone is not enough. There must also be a clear method for determining the extent of each actor's responsibility.

The Multi-Tiered Accountability Framework proposed in this paper attempts to address this gap by connecting technical accountability, actor accountability and responsibility allocation. It recognises that accountability should exist throughout the AI lifecycle rather than only after harm has occurred.

The framework does not argue that autonomous AI should be stopped or that every unexpected AI decision should result in human liability. Instead, it argues that greater autonomy should not mean weaker accountability. Responsibility should depend on the circumstances of each case and should be connected to an actor's contribution, control, ability to foresee the risk and ability to prevent or reduce the harm.

Technology is evolving, and law has to evolve with it. The challenge for the law is therefore not to prevent AI from becoming more autonomous, but to ensure that autonomy does not create a space where responsibility disappears. A system may be autonomous in its decision-making, but the legal system must still be able to answer the question: When something goes wrong, who was in the best position to prevent it, and why?