# The Future of Legal Responsibility in Human–AI Collaboration: Revisiting Fault, Intent, and Causation in Hybrid Decision Structures

**1. Georgios Nikolaidis: Department of International and European Studies, University of Piraeus, Piraeus, Greece**
**2. Andrei Ionescu*: Department of Private Law, University of Bucharest, Bucharest, Romania**

*Correspondence: e-mail: andrei.ionescu@unibuc.ro

**Abstract**

The rapid integration of artificial intelligence into contemporary decision-making processes has fundamentally altered the structure of agency, accountability, and risk within modern legal systems. Human–AI collaboration now characterizes critical domains such as healthcare, finance, transportation, governance, and criminal justice, producing decisions through complex interactions between human judgment and algorithmic inference. This article examines how these hybrid decision structures destabilize the classical foundations of legal responsibility, particularly the doctrines of fault, intent, and causation. Employing a narrative review methodology grounded in descriptive–analytical inquiry, the study synthesizes interdisciplinary scholarship from law, philosophy of action, AI governance, and socio-technical systems theory to reconstruct the conceptual architecture of responsibility under conditions of distributed cognition. The analysis demonstrates that traditional anthropocentric models of responsibility—premised on individual agency, linear causation, and coherent intentionality—are increasingly inadequate for explaining harm and allocating accountability in algorithmically mediated environments. The article proposes a systemic reorientation of legal responsibility, emphasizing shared and layered accountability, institutional governance, and risk-based causation frameworks. By reframing responsibility as a property of socio-technical systems rather than isolated individuals, the study offers a coherent theoretical foundation for adapting liability regimes to the realities of human–AI collaboration. The findings suggest that the future legitimacy and effectiveness of legal systems depend on their capacity to evolve beyond event-based blame toward governance-centered models capable of sustaining accountability amid technological complexity.

**Keywords:** Human–AI collaboration; legal responsibility; fault and intent; causation theory; hybrid decision-making; algorithmic governance; distributed agency; risk governance

Received: date: 20 August 2023
Revised: date: 14 September 2023
Accepted: date: 27 September 2023
Published: date: 01 October 2023



**Citation**: Nikolaidis, G., & Ionescu, A. (2023). The Future of Legal Responsibility in Human–AI Collaboration: Revisiting Fault, Intent, and Causation in Hybrid Decision Structures. *Legal Studies in Digital Age*, 2(4), 48-60.

## 1. Introduction

The rapid integration of artificial intelligence into decision-making processes has fundamentally transformed the architecture of action, responsibility, and authority across contemporary social, economic, and legal systems. What was once characterized by relatively clear lines between human agency and technological assistance has evolved into complex socio-technical

assemblages in which decisions are increasingly co-produced by humans and machine systems. In healthcare, for example, diagnostic recommendations are now routinely generated through algorithmic models trained on vast clinical datasets, reshaping professional judgment and institutional workflows (Ahsen et al., 2023). Similar transformations are observable in finance, governance, transportation, and environmental management, where predictive systems guide strategic choices, risk assessments, and regulatory interventions (Mühling, 2023). These developments do not merely introduce new tools into existing decision structures; rather, they reconfigure the very ontology of decision-making by distributing cognition, authority, and causal influence across human and non-human actors. As AI systems increasingly participate in epistemic processes, producing interpretations, classifications, and predictions that shape outcomes, the classical image of a sovereign human decision-maker operating over passive instruments becomes progressively untenable.

The evolution from automation toward collaborative intelligence represents a critical conceptual shift in the role of machines within institutional and legal frameworks. Early automation technologies were designed primarily to execute predefined tasks under strict human control, reinforcing a hierarchical model in which humans retained ultimate authority and responsibility. Contemporary AI systems, by contrast, engage in adaptive learning, pattern recognition, and probabilistic inference, allowing them to generate novel outputs that were neither explicitly programmed nor fully foreseeable by their creators (Yang, 2022). In high-stakes environments such as critical care medicine, clinicians increasingly interact with AI not as mere tools but as cognitive partners whose recommendations shape clinical reasoning and therapeutic decisions (Bienefeld et al., 2023). This shift toward collaborative intelligence blurs the boundaries between human judgment and machine inference, giving rise to hybrid decision structures in which agency is distributed across interconnected human–machine networks (Veitch et al., 2022). Such hybridity challenges foundational assumptions of legal doctrine that presuppose identifiable human authorship behind each legally relevant act.

The epistemic status of AI within these hybrid structures further complicates the traditional legal understanding of action and responsibility. AI systems are no longer confined to instrumental functions; they increasingly operate as epistemic agents that transform raw data into knowledge claims that influence human belief formation and institutional behavior (Jing-jing et al., 2023). In clinical contexts, for instance, machine learning models provide diagnostic probabilities that clinicians incorporate into treatment planning, effectively reshaping professional expertise and redefining standards of care (Hryciw, Seely, et al., 2023). At the same time, the opacity of many algorithmic systems undermines traditional criteria for attributing knowledge, intent, and fault, as neither users nor developers may fully comprehend the internal reasoning processes of complex models (Lang et al., 2023). This epistemic transformation destabilizes established doctrines of foreseeability and control, which have historically served as cornerstones of legal responsibility.

Parallel to these epistemic shifts, the expansion of algorithmic governance has embedded AI deeply within regulatory, administrative, and political decision-making processes. Governments increasingly rely on automated systems for welfare allocation, risk profiling, immigration control, and public resource management, thereby integrating algorithmic outputs into the core functions of state authority (Brand, 2022). Such developments amplify the normative stakes of hybrid decision-making, as algorithmic interventions now directly affect rights, obligations, and life opportunities of individuals and communities (Leslie et al., 2021). The freedom to conduct a business, democratic participation, and access to essential services are progressively mediated by algorithmic infrastructures, creating new forms of power asymmetry and accountability gaps (Neves, 2023). Within this evolving landscape, legal responsibility can no longer be examined solely through the lens of individual wrongdoing or negligence but must be situated within broader socio-technical systems of governance.

These transformations precipitate a profound crisis for traditional models of legal responsibility. Classical legal doctrine is grounded in a subject–object dichotomy that presupposes autonomous human actors exercising control over inert tools. Within this framework, responsibility is attributed by identifying the human subject whose voluntary action caused a legally relevant outcome. Hybrid decision structures disrupt this dichotomy by introducing machine systems that contribute substantively to decision formation, outcome production, and risk distribution. The notion that AI merely extends human will is increasingly implausible in contexts where machine outputs shape choices in ways that even expert users cannot fully anticipate (Sio & Mecacci, 2021). Responsibility thus becomes diffuse, fragmented across designers, developers, operators, users, institutions,

and the AI systems themselves, generating what scholars have described as responsibility gaps (Lang et al., 2023). These gaps undermine the coherence of fault-based liability regimes that depend on traceable lines of causation and identifiable agents of wrongdoing.

The inadequacy of anthropocentric liability frameworks becomes particularly evident when legal analysis confronts the layered complexity of human–AI collaboration. Traditional doctrines of negligence, intent, and causation were constructed for environments in which human cognition and action served as the primary sources of legal meaning. In hybrid systems, however, machine-generated outputs influence human perception, decision priorities, and risk evaluations, thereby co-authoring outcomes in ways that challenge simple attribution models (Smith & Fotheringham, 2020). In medical practice, for instance, liability cannot be fully explained by reference to either physician negligence or product defect when clinical decisions emerge from continuous interaction between clinician expertise and algorithmic recommendation (Carter et al., 2020). Similar dilemmas arise in autonomous transportation, algorithmic trading, and predictive policing, where outcomes are shaped by feedback loops between human supervision and machine autonomy (Gillespie, 2022). Existing legal frameworks struggle to capture these dynamics, exposing the need for conceptual reconstruction.

Against this backdrop, the present study addresses a central problem: whether the foundational doctrines of fault, intent, and causation can remain viable in legal systems increasingly structured around hybrid human–AI decision-making. When cognition, agency, and control are distributed across socio-technical networks, the classical architecture of responsibility is destabilized. This raises urgent normative and doctrinal questions. Can fault survive when harmful outcomes result from probabilistic models trained on vast datasets beyond direct human comprehension? How should intent be interpreted when decisions reflect the convergence of human objectives and machine-generated inferences? What constitutes causation in environments where outcomes emerge from complex, non-linear interactions between human actors and adaptive algorithms? These questions are no longer theoretical abstractions; they confront courts, regulators, and institutions across multiple jurisdictions (Arsenijević, 2023). Furthermore, they expose the limitations of legal systems that continue to rely on conceptual categories forged in pre-digital contexts.

The issue of responsibility becomes even more intricate when decisions are co-produced across multiple institutional layers. Designers encode normative assumptions into system architectures (Караджоянни, 2021), organizations shape deployment contexts and governance policies (Ardichvili, 2022), and end users integrate algorithmic outputs into practical decision-making under conditions of uncertainty (Hajnal, 2021). Harmful outcomes may therefore reflect the cumulative effect of design choices, organizational incentives, regulatory omissions, and individual actions. Assigning responsibility within such ecosystems requires moving beyond linear models of blame toward more systemic frameworks of accountability (Gillespie, 2022). Yet existing liability regimes remain predominantly individualistic, focusing on discrete acts of misconduct rather than structural patterns of risk production.

This study is situated within this unfolding legal and technological transformation and seeks to contribute to the reconstruction of responsibility doctrine for the era of human–AI collaboration. The research examines how fault, intent, and causation must be reinterpreted when decision-making is no longer exclusively human but distributed across hybrid cognitive systems. It explores the normative consequences of shifting from individualistic models of liability toward more systemic and governance-oriented approaches that recognize the socio-technical nature of contemporary harm production (Lang et al., 2023). In doing so, the article interrogates the conceptual boundaries of legal agency, the epistemic foundations of liability, and the institutional conditions under which accountability can be meaningfully sustained in algorithmically mediated societies.

Methodologically, the study adopts a narrative review design grounded in a descriptive–analytical approach. This orientation enables comprehensive synthesis of interdisciplinary scholarship spanning doctrinal law, philosophy of mind, AI ethics, technology governance, and jurisprudence. By integrating insights from legal theory, cognitive science, and socio-technical systems analysis, the research constructs a conceptual map of responsibility under hybrid conditions. The narrative review framework allows the examination of evolving conceptual trends, normative debates, and regulatory responses across diverse jurisdictions and institutional settings. Sources include legal judgments, statutory frameworks, theoretical treatises, policy documents, and empirical studies addressing the ethical, legal, and social implications of AI integration (Hill et al., 2023).

This integrative approach facilitates critical evaluation of existing responsibility models while identifying conceptual gaps and future directions for legal development.

The descriptive–analytical method further enables systematic articulation of how legal doctrines are being strained and transformed by technological change. Rather than proposing a single normative solution, the study traces the shifting contours of responsibility discourse, highlighting competing frameworks and emerging consensus where applicable. It situates legal responsibility within the broader political economy of algorithmic governance, acknowledging the role of institutional incentives, power structures, and regulatory design in shaping accountability outcomes (Neves, 2023). By foregrounding the structural conditions under which human–AI collaboration operates, the analysis avoids reductive individualism and advances a more nuanced understanding of how responsibility is constructed, negotiated, and enforced in contemporary legal systems.

Ultimately, the necessity of revisiting foundational doctrines of responsibility arises from the recognition that human–AI collaboration is not a peripheral phenomenon but a defining feature of modern governance and social organization. As AI systems continue to permeate decision-making infrastructures, the legitimacy of legal systems depends on their capacity to adapt core concepts of fault, intent, and causation to hybrid realities. Without such adaptation, legal doctrine risks becoming increasingly detached from the socio-technical conditions it purports to regulate, undermining both normative coherence and practical effectiveness. This study therefore undertakes a comprehensive re-examination of responsibility in the age of collaborative intelligence, seeking to lay the conceptual groundwork for a new legal architecture capable of sustaining accountability in algorithmically mediated societies.

## 2. Theoretical Foundations of Legal Responsibility in the Age of Human–AI Collaboration

The transformation of decision-making through artificial intelligence necessitates a fundamental reconstruction of the theoretical foundations of legal responsibility. Classical legal doctrine was developed within an epistemic environment where human cognition, intention, and control were assumed to be the primary drivers of legally relevant action. Under such conditions, the architecture of responsibility rested on relatively stable concepts of fault, intent, and causation that presupposed identifiable human agents acting through transparent chains of reasoning and action. In contemporary socio-technical systems, however, these assumptions are increasingly destabilized. AI systems now participate in the production of knowledge, the framing of choices, and the execution of outcomes, thereby redistributing cognitive and causal influence across human and non-human components (Ahsen et al., 2023). This redistribution of agency challenges the coherence of inherited responsibility models and compels legal theory to reconsider the conceptual foundations upon which accountability is constructed.

Within classical responsibility frameworks, fault-based liability has long served as the dominant paradigm for attributing legal responsibility. Rooted in notions of negligence and culpability, fault-based models presume that individuals possess sufficient control over their actions and sufficient knowledge of foreseeable consequences to justify moral and legal blame. This framework remains central to tort law and professional liability regimes, particularly in high-stakes domains such as medicine and engineering (Smith & Fotheringham, 2020). Yet in human–AI collaboration, the foreseeability and controllability that fault-based liability requires are progressively eroded by the opacity and adaptive learning characteristics of machine systems. Clinical practitioners, for instance, increasingly rely on algorithmic recommendations that they cannot fully audit or replicate, complicating determinations of whether adverse outcomes result from human negligence, system error, or emergent interactions between the two (Hryciw, Seely, et al., 2023). As a result, fault-based liability risks either over-attributing blame to individual users or under-attributing responsibility where systemic risks are the true source of harm (Lang et al., 2023).

Strict liability offers an alternative approach that circumvents the epistemic challenges of proving fault by imposing responsibility regardless of intent or negligence. Historically justified in contexts involving abnormally dangerous activities or defective products, strict liability prioritizes risk distribution and victim compensation over moral blameworthiness. In the context of AI, strict liability has gained attention as a potential mechanism for addressing harms arising from complex algorithmic systems whose internal operations defy transparent analysis (Putranti & Anggraeny, 2022). However, while strict liability simplifies adjudication, it also risks obscuring the layered contributions of designers, deployers, and institutional actors who shape system behavior through choices about training data, deployment environments, and governance structures (Brand,

2022). Without careful calibration, strict liability may entrench corporate concentration of risk while failing to incentivize robust governance and ethical system design.

Intentional wrongdoing represents the third pillar of classical responsibility, grounded in the premise that legal culpability intensifies when harmful outcomes are produced through deliberate and conscious decision-making. Criminal law, in particular, relies heavily on the doctrine of mens rea to differentiate between degrees of culpability based on subjective mental states. Yet the attribution of intent becomes profoundly problematic in hybrid human–AI contexts. When decisions reflect the convergence of human objectives and machine-generated inferences, disentangling the locus of intent becomes conceptually unstable (Sio & Mecacci, 2021). Algorithmic systems may generate outputs that no human explicitly intended, while human actors may act upon recommendations whose internal logic they do not fully comprehend (Yang, 2022). This fragmentation of intentionality destabilizes traditional mens rea categories and exposes the inadequacy of legal models that presuppose coherent, unified mental states behind each legally significant act.

These challenges invite deeper engagement with the philosophy of action and agency, where longstanding debates about intentionality, autonomy, and control provide conceptual tools for reconstructing responsibility under hybrid conditions. Human intentionality has traditionally been understood as the capacity to act for reasons, guided by beliefs and desires that render behavior intelligible and normatively assessable. Legal responsibility presupposes this model of agency, treating individuals as rational authors of their actions. However, contemporary human–AI collaboration introduces new forms of mediated intentionality, in which algorithmic systems reshape the informational environment within which human intentions are formed (Jing-jing et al., 2023). Decision-makers increasingly operate within epistemic scaffolds constructed by machine systems, thereby complicating the attribution of autonomous intention.

Agency, autonomy, and control are likewise destabilized in socio-technical systems. Autonomy presupposes the capacity for self-governance, while control implies the ability to guide outcomes in accordance with one's intentions. In hybrid decision environments, both capacities are distributed across networks of human and machine actors. For example, clinicians retain formal authority over medical decisions, yet their choices are often constrained and guided by algorithmic risk scores and predictive models (Bienefeld et al., 2023). This erosion of unilateral control undermines the foundational premise of many responsibility doctrines that equate agency with individual mastery over action. Scholars have therefore called for reconceptualizing agency as an emergent property of socio-technical systems rather than a discrete attribute of isolated individuals (Gillespie, 2022).

The concept of collective and distributed agency further illuminates the transformation of responsibility in hybrid systems. Collective agency recognizes that groups, organizations, and institutions can function as intentional actors whose coordinated actions produce outcomes irreducible to individual contributions. Distributed agency extends this insight to socio-technical networks, where humans and machines jointly produce decisions through complex interactions (Veitch et al., 2022). In such contexts, responsibility cannot be meaningfully located in a single node of the system but must be understood as emerging from the structure and governance of the network as a whole (Hajnal, 2021). This shift challenges legal doctrines that continue to privilege individual blame over systemic accountability.

The question of whether AI itself should be regarded as a new legal actor further complicates the reconstruction of responsibility. The traditional tool paradigm treats AI as an inert instrument through which human will is executed, thereby locating responsibility exclusively in human actors. Yet as AI systems acquire increasing autonomy, adaptability, and influence over outcomes, this paradigm becomes increasingly inadequate (Eck & Agbeko, 2023). Some scholars have argued that AI systems exhibit functional characteristics of agency, including goal-directed behavior and context-sensitive adaptation, that warrant reconsideration of their legal status (Arsenijević, 2023). Others caution that attributing legal agency to machines risks diluting human accountability and obscuring the political and economic forces shaping AI deployment (Leslie et al., 2021).

The distinction between moral and legal agency is central to this debate. Moral agency presupposes the capacity for normative understanding and ethical deliberation, qualities that current AI systems do not possess. Legal agency, by contrast, is a normative construct designed to allocate rights and obligations within institutional frameworks. Corporations, for instance, are legal persons despite lacking moral consciousness. This distinction opens the possibility of limited forms of AI legal agency oriented toward responsibility allocation rather than moral blame (Eck & Agbeko, 2023). Nevertheless, extending legal

personhood to AI remains highly contested, as it risks legitimizing the displacement of human accountability onto technological artifacts.

Algorithmic opacity and epistemic gaps further complicate any attempt to integrate AI into responsibility frameworks. Many machine learning systems operate as "black boxes" whose internal processes are inaccessible even to their developers (Lang et al., 2023). This opacity undermines traditional evidentiary standards and frustrates efforts to trace causal pathways between actions and outcomes (Smith & Fotheringham, 2020). Explainable AI initiatives seek to mitigate these problems by enhancing transparency and interpretability, yet significant epistemic uncertainties persist (Yang, 2022). Legal systems must therefore grapple with responsibility in environments where knowledge of system behavior is inherently incomplete.

Hybrid cognition and distributed decision-making represent the culmination of these theoretical transformations. Contemporary socio-technical systems integrate human judgment with algorithmic processing in continuous feedback loops, producing decisions that cannot be attributed to either component in isolation (Ahsen et al., 2023). Cognitive offloading, whereby humans delegate memory, calculation, and pattern recognition tasks to machines, reshapes the structure of human cognition itself (Mohanasundari et al., 2023). While such delegation enhances efficiency and accuracy in many domains, it also redistributes epistemic authority and blurs the boundary between human and machine agency (Hryciw, Seely, et al., 2023). Responsibility diffusion emerges as an inevitable consequence, as the increasing complexity of decision networks makes it difficult to isolate individual contributions to harmful outcomes (Sio & Mecacci, 2021).

In these environments, responsibility must be reconceptualized as a property of socio-technical systems rather than a feature of isolated agents. Governance structures, institutional incentives, regulatory frameworks, and system design choices collectively shape risk production and harm distribution (Brand, 2022). Ethical frameworks for responsible AI emphasize the need for human-centered design, accountability mechanisms, and continuous oversight to counteract the centrifugal forces of responsibility diffusion (Sqalli et al., 2023). Yet such frameworks remain underdeveloped in law, which continues to rely on doctrinal categories forged in an era of human-exclusive agency.

The theoretical foundations of legal responsibility are therefore undergoing a profound transformation. As human–AI collaboration becomes the dominant mode of decision-making, the conceptual architecture of responsibility must evolve to reflect the distributed, opaque, and emergent nature of contemporary action. This reconstruction does not entail abandoning the core normative commitments of responsibility—fairness, accountability, and justice—but rather rearticulating them within a new socio-technical ontology. Only through such conceptual renewal can legal systems sustain legitimacy and effectiveness in the age of hybrid cognition.

## 3. Reconstructing Fault and Intent in Human–AI Systems

The accelerating integration of artificial intelligence into core decision-making processes compels a fundamental rethinking of how legal systems conceptualize fault and intent. Traditional doctrines of negligence and mens rea were developed in socio-technical contexts where human cognition, judgment, and control were the dominant sources of legally relevant conduct. In AI-mediated environments, however, decisions emerge from continuous interaction between human actors and algorithmic systems whose internal processes are often opaque, probabilistic, and adaptive. In healthcare diagnostics, for example, clinicians increasingly rely on machine-generated risk scores and predictive models that influence clinical reasoning in ways neither fully transparent nor entirely controllable (Ahsen et al., 2023). Similar dynamics characterize finance, transportation, and governance, where algorithmic outputs shape professional and institutional conduct (Brand, 2022). These transformations destabilize the classical foundations of fault and intent by disrupting the assumptions of foreseeability, control, and unified intentionality upon which legal responsibility has long depended.

Fault in classical negligence doctrine presupposes that actors can reasonably foresee the consequences of their conduct and adjust their behavior accordingly. Yet in algorithmic environments, foreseeability is structurally compromised by the statistical and non-deterministic nature of machine learning systems. Models trained on large datasets generate probabilistic predictions that evolve over time through continuous learning, often producing outputs that even their developers cannot precisely anticipate (Yang, 2022). In clinical decision-making, this uncertainty complicates determinations of whether adverse outcomes

result from negligent reliance on AI recommendations or from the inherent unpredictability of complex models (Smith & Fotheringham, 2020). The concept of negligence must therefore be recalibrated to account for the epistemic limits of both human users and system designers operating within algorithmic uncertainty (Lang et al., 2023).

Foreseeability under machine learning conditions cannot be equated with the traditional expectation that reasonable actors can predict specific outcomes. Instead, foreseeability increasingly concerns awareness of systemic risks, limitations of model reliability, and potential failure modes of algorithmic systems. Medical practitioners who deploy AI tools, for instance, may not foresee the precise diagnostic error that leads to patient harm, but they can foresee that algorithmic bias, data quality issues, and model drift create non-trivial risks of error (Carter et al., 2020). Similarly, organizations deploying automated systems for resource allocation or risk assessment must anticipate that such systems can generate discriminatory or erroneous outcomes even when functioning as designed (Leslie et al., 2021). Fault thus shifts from individual misjudgment toward failures of institutional risk management and governance.

This shift also destabilizes the traditional "reasonable person" standard that underpins negligence law. The reasonable person standard presumes a generic human actor exercising ordinary prudence under given circumstances. In AI-mediated contexts, however, decisions are no longer made by unaided individuals but by hybrid human–machine ensembles. A physician relying on clinical AI, an engineer supervising autonomous systems, or a financial analyst using predictive algorithms operates within epistemic conditions radically different from those contemplated by classical negligence doctrine (Hryciw, Seely, et al., 2023). Some scholars therefore argue for the emergence of a "reasonable designer" or "reasonable operator" standard, calibrated to the technical expertise, informational access, and institutional role of actors involved in AI deployment (Putranti & Anggraeny, 2022). Under such standards, fault is assessed not by reference to ordinary human judgment but by evaluating whether actors fulfilled their professional and organizational duties to design, validate, monitor, and govern algorithmic systems responsibly (Sqalli et al., 2023).

The reconstruction of intent in human–AI systems presents even deeper conceptual challenges. Classical mens rea doctrine presupposes a coherent mental state that accompanies the commission of a legally relevant act. In hybrid decision-making environments, however, intentionality becomes fragmented across human objectives, algorithmic inferences, and institutional constraints. Algorithmic outputs often influence decisions in ways that no human actor explicitly intends. A clinician may intend to follow best medical practice, yet act on an AI recommendation that reflects hidden correlations or biased training data, producing unintended harm (Hryciw, Fortin, et al., 2023). Likewise, a trader may deploy an algorithmic trading system to maximize profit without intending the specific market manipulation effects produced by high-frequency strategies (Santos, 2023). The divergence between human intent and system output destabilizes the traditional alignment between mental state and harmful consequence upon which criminal and civil liability depend.

This fragmentation of mens rea undermines the coherence of intent-based responsibility. When harmful outcomes emerge from the interaction between human oversight and machine autonomy, attributing intent to any single actor becomes conceptually strained. Scholars have described this as one of the core "responsibility gaps" generated by AI, in which neither the human user nor the system designer can be said to possess full intentional control over outcomes (Sio & Mecacci, 2021). At the same time, legal systems cannot simply abandon the notion of intent without undermining the normative foundations of culpability. The challenge, therefore, lies in developing doctrines that recognize degrees and distributions of intentional contribution across socio-technical networks (Lang et al., 2023).

Delegated decision authority further complicates intent attribution. In many institutional contexts, humans formally delegate decision-making tasks to AI systems while retaining ultimate legal authority over outcomes. In medicine, clinicians remain legally responsible for treatment decisions even when they follow algorithmic recommendations (Smith & Fotheringham, 2020). In financial markets, traders and firms retain liability for algorithmic trading activities despite limited real-time control over system behavior (Hasan, 2022). This delegation creates asymmetries between legal responsibility and practical control, as actors are held accountable for outcomes that emerge from processes they cannot fully supervise. Such asymmetries expose the inadequacy of existing intent doctrines and highlight the need for new models of distributed culpability.

In response to these challenges, legal scholarship increasingly explores shared and layered responsibility frameworks that distribute accountability across multiple actors and institutional levels. Designers shape system behavior through choices about data, model architecture, and training objectives (Караджоянни, 2021). Deployers determine the contexts and purposes for which systems are used, influencing risk profiles and potential harms (Brand, 2022). Operators interact with AI outputs in real time, integrating them into concrete decisions under conditions of uncertainty (Veitch et al., 2022). End users experience the downstream consequences of these decisions, often without knowledge of the underlying processes (Hill et al., 2023). Shared responsibility models seek to capture this distributed causality by allocating duties and liabilities according to each actor's role in the socio-technical system (Gillespie, 2022).

Organizational liability plays a central role in these emerging frameworks. Corporations, hospitals, financial institutions, and government agencies function as collective actors whose policies, incentives, and governance structures shape the behavior of human–AI systems. Ethical failures in AI deployment frequently reflect organizational decisions about cost, efficiency, and risk tolerance rather than isolated acts of individual negligence (Ardichvili, 2022). Recognizing this, scholars advocate strengthening institutional accountability mechanisms, including mandatory risk assessments, transparency obligations, and internal oversight structures (Neves, 2023). Organizational liability thus becomes a crucial bridge between individual responsibility and systemic governance in AI-mediated environments.

Risk governance frameworks provide an additional layer of conceptual support for reconstructing fault and intent. Rather than focusing solely on post hoc blame, risk governance emphasizes anticipatory regulation, continuous monitoring, and adaptive control of emerging technologies (Sabt & Farooqui, 2023). In the AI context, this approach aligns with calls for human-centered design, ethical impact assessments, and ongoing system auditing to mitigate foreseeable harms (Sqalli et al., 2023). By embedding responsibility within governance processes, legal systems can better accommodate the fluid, evolving nature of AI systems while preserving normative commitments to accountability and justice.

The practical implications of these conceptual shifts become especially visible in case-based contexts. Autonomous vehicles exemplify the breakdown of traditional fault and intent doctrines. When a self-driving car causes an accident, responsibility may involve the vehicle's manufacturer, software developers, data providers, fleet operators, and the human occupant. The human driver may not have intended or even been capable of preventing the specific malfunction, while designers may not have foreseen the precise environmental conditions that triggered the failure (Kalra, 2022). Assigning fault under such conditions requires moving beyond simplistic negligence models toward layered responsibility structures that reflect the distributed nature of causation and control (Putranti & Anggraeny, 2022).

Algorithmic trading presents parallel challenges. High-frequency trading systems execute thousands of transactions per second based on predictive models that adapt to market conditions in real time. Traders and firms may intend to maximize profit, yet the cumulative effects of algorithmic strategies can destabilize markets, amplify volatility, or produce systemic risks that no individual actor explicitly intended (Hasan, 2022). Traditional intent-based liability struggles to capture such phenomena, as harmful outcomes arise from emergent interactions among multiple autonomous systems. Regulatory responses increasingly emphasize organizational accountability, risk controls, and systemic oversight rather than individual fault (Brand, 2022).

AI in medicine and criminal justice further illustrates the urgency of reconstructing fault and intent. In clinical contexts, AI diagnostic tools influence treatment decisions, shaping patient outcomes in ways that neither clinicians nor patients fully understand (Ahsen et al., 2023). When harm occurs, courts must navigate the tension between professional negligence standards and the realities of algorithmic mediation (Smith & Fotheringham, 2020). In criminal justice, predictive policing and risk assessment algorithms shape bail, sentencing, and parole decisions, raising profound questions about intent, discrimination, and accountability (Leslie et al., 2021). Judges and prosecutors may rely on algorithmic scores without intending the biased or unjust outcomes those systems sometimes produce (Jing-jing et al., 2023). These cases expose the inadequacy of traditional responsibility doctrines and underscore the necessity of systemic, governance-oriented reforms.

Across these domains, fault and intent are no longer attributes of isolated individuals but emergent properties of complex socio-technical systems. Reconstructing these doctrines requires abandoning purely anthropocentric models of responsibility

and embracing frameworks that recognize distributed cognition, shared agency, and institutional governance as central features of contemporary decision-making. Only through such conceptual renewal can legal systems preserve accountability while accommodating the transformative impact of human–AI collaboration.

## 4. Rethinking Causation in Algorithmic and Hybrid Decision Contexts

The doctrine of causation occupies a central position in legal responsibility, serving as the conceptual bridge between conduct and consequence. Classical causation theory, developed in an era of predominantly human-driven action, presupposes linear chains of events, identifiable agents, and relatively transparent mechanisms of influence. In algorithmic and hybrid decision environments, however, these assumptions no longer hold. AI systems operate through complex, probabilistic, and adaptive processes that reshape how outcomes emerge and how responsibility can be meaningfully attributed. In medical diagnostics, for instance, predictive models generate probabilistic risk assessments that guide clinical decisions, yet the precise pathway from algorithmic inference to patient outcome remains opaque and contingent (Ahsen et al., 2023). Similar dynamics arise in finance, governance, transportation, and environmental management, where algorithmic interventions mediate causal processes in ways that classical doctrine struggles to accommodate (Mühling, 2023). Rethinking causation is therefore indispensable to preserving the coherence of legal responsibility in the age of human–AI collaboration.

Classical causation doctrines rely heavily on the but-for test, which asks whether the harm would have occurred in the absence of the defendant's conduct. This test presupposes relatively simple causal structures in which individual actions can be isolated and their effects traced through direct sequences of events. In hybrid systems, however, harms often result from the convergence of multiple algorithmic and human factors, making counterfactual analysis deeply problematic. A clinician's decision influenced by an AI recommendation cannot easily be separated from the system's training data, model architecture, institutional protocols, and broader clinical environment (Smith & Fotheringham, 2020). Removing any single element from this socio-technical network may not yield a clear alternative outcome, rendering the but-for test conceptually unstable (Lang et al., 2023). As AI systems increasingly participate in causal processes, traditional counterfactual reasoning loses much of its explanatory power.

Proximate cause doctrine, which limits liability to harms that bear a sufficiently close connection to the defendant's conduct, encounters similar difficulties. Proximate cause presumes a normative judgment about which causal links are legally significant, grounded in notions of foreseeability and directness. In algorithmic systems, however, causal pathways are frequently indirect, mediated through layers of data processing, feedback mechanisms, and adaptive learning processes (Yang, 2022). A biased dataset introduced at the design stage may influence model behavior months later, producing discriminatory outcomes in a different institutional context (Leslie et al., 2021). Determining whether such distant and mediated influences satisfy proximate cause requirements exposes the limitations of doctrines crafted for more immediate and tangible forms of causation.

The challenge intensifies in the presence of multi-actor causal chains. Hybrid decision environments typically involve designers, developers, data providers, deployers, operators, and end users, each contributing to the final outcome. In autonomous vehicle systems, for example, sensor manufacturers, software engineers, fleet operators, and human occupants collectively shape the vehicle's behavior (Kalra, 2022). When an accident occurs, the harm reflects the interaction of these multiple actors rather than the isolated conduct of any single participant. Classical causation frameworks, oriented toward individual defendants and discrete actions, struggle to allocate responsibility within such distributed networks (Gillespie, 2022). The resulting gaps undermine both the fairness and the deterrent function of liability regimes.

Algorithmic causality introduces additional complexity by departing from deterministic models of cause and effect. Machine learning systems operate through statistical inference rather than rule-based logic, generating outputs that reflect probabilistic correlations rather than necessary consequences (Yang, 2022). In healthcare, predictive models may identify patients as high risk based on patterns in large datasets, yet no single factor can be said to deterministically cause the predicted outcome (Ahsen et al., 2023). When decisions informed by such models result in harm, causation cannot be understood as a direct link between specific inputs and specific outputs but as the product of statistical processes embedded in dynamic systems (Hryciw, Seely,

et al., 2023). This shift from deterministic to probabilistic causation challenges legal doctrines that presuppose identifiable causal mechanisms.

Statistical and probabilistic causation further complicate attribution by introducing uncertainty into the causal analysis. Courts have traditionally been reluctant to base liability on probabilistic evidence, preferring concrete proof of causal connection. Yet in algorithmic environments, probabilistic reasoning is often the only available mode of explanation. For example, risk assessment tools in criminal justice predict recidivism probabilities that influence sentencing and parole decisions (Jing-jing et al., 2023). When such predictions contribute to unjust or discriminatory outcomes, causation resides not in a single erroneous decision but in the cumulative effect of probabilistic inferences operating within institutional frameworks (Leslie et al., 2021). Legal doctrine must therefore evolve to recognize probabilistic causation as a legitimate basis for responsibility.

Emergent outcomes represent perhaps the most profound challenge to classical causation. Hybrid systems exhibit emergent behavior when interactions among components produce results that cannot be predicted from the properties of individual elements. Algorithmic trading systems, for instance, may collectively generate market instability through feedback loops and strategic interactions that no single trader or system designer intended (Hasan, 2022). These emergent phenomena defy linear causal explanation and expose the inadequacy of doctrines that seek to isolate individual causes in complex systems (Sio & Mecacci, 2021). Responsibility in such contexts must grapple with system-level properties rather than discrete actions.

Hybrid causal architectures crystallize these challenges by integrating human input and algorithmic processing into continuous decision cycles. Human actors provide goals, constraints, and oversight, while AI systems generate predictions, recommendations, and automated actions. In medical practice, clinicians interpret algorithmic outputs, adjust treatment plans, and feed new data back into learning systems (Bienefeld et al., 2023). These interactions form feedback loops in which each component continuously influences the others, producing outcomes that reflect the system as a whole rather than any single contributor (Veitch et al., 2022). Classical causation, with its emphasis on linear sequences, cannot adequately capture these recursive dynamics.

Feedback loops intensify the problem of responsibility dilution by dispersing causal influence across multiple system components. In predictive policing, for example, algorithmic forecasts guide law enforcement deployment, generating new data that reinforces existing patterns of surveillance and bias (Leslie et al., 2021). The resulting outcomes emerge from cyclical interactions between human decision-makers and algorithmic systems, making it difficult to locate responsibility at any particular point in the process (Gillespie, 2022). As causal influence becomes increasingly diffuse, the risk of accountability erosion grows, threatening the legitimacy of legal responsibility frameworks.

In response to these challenges, scholars and policymakers are exploring new causation frameworks better suited to probabilistic and opaque systems. Risk-based causation shifts the focus from proving specific causal chains to evaluating whether actors created or exacerbated unreasonable risks of harm. This approach aligns with emerging regulatory strategies that emphasize precaution, impact assessment, and continuous monitoring of AI systems (Sabt & Farooqui, 2023). By foregrounding risk creation rather than direct causation, legal doctrine can better accommodate the uncertainties inherent in algorithmic decision-making.

Systemic causation offers another promising direction by treating socio-technical systems as the primary locus of causal analysis. Under this framework, responsibility is attributed based on an actor's role in shaping system behavior rather than on isolated actions. Organizations that design, deploy, and govern AI systems bear responsibility for systemic harms that arise from predictable patterns of interaction within those systems (Brand, 2022). This perspective recognizes that many AI-related harms reflect structural features of institutions and markets rather than individual misconduct (Ardichvili, 2022). Systemic causation thus provides a conceptual bridge between individual liability and institutional accountability.

Governance-based attribution models further extend this systemic orientation by embedding responsibility within regulatory and organizational structures. These models emphasize the importance of transparency, oversight, and participatory governance in managing the causal complexity of AI systems (Neves, 2023). Ethical frameworks for responsible AI advocate for clear lines of accountability across the AI lifecycle, from design and development to deployment and monitoring (Sqalli et al., 2023).

By integrating causation analysis into governance processes, legal systems can move beyond reactive adjudication toward proactive harm prevention and accountability.

The rethinking of causation in algorithmic and hybrid decision contexts therefore demands a paradigmatic shift in legal theory. Classical doctrines rooted in linear, deterministic models of cause and effect must give way to frameworks capable of addressing probabilistic reasoning, emergent behavior, and distributed agency. This transformation does not negate the normative foundations of responsibility but rather rearticulates them within a socio-technical ontology attuned to the realities of contemporary decision-making. Only by embracing new causation models can legal systems sustain accountability, fairness, and legitimacy in an era defined by human–AI collaboration.

## 5. Conclusion

The transformation of decision-making through artificial intelligence represents one of the most profound challenges contemporary legal systems have ever confronted. Human–AI collaboration is no longer an emerging phenomenon; it has become a defining structural feature of modern governance, professional practice, and economic life. This article has demonstrated that the classical architecture of legal responsibility—built upon stable assumptions of human agency, linear causation, and transparent intention—cannot adequately accommodate the realities of hybrid decision environments. Fault, intent, and causation, which once functioned as coherent and reliable pillars of accountability, now operate within epistemic and organizational conditions that fundamentally alter their meaning and effectiveness.

At the core of this transformation lies the collapse of the traditional subject–object model of responsibility. AI systems do not simply execute human will; they shape perception, structure choices, and generate epistemic content that influences outcomes in decisive ways. As a result, responsibility can no longer be conceptualized as a property of isolated individuals acting through passive tools. It emerges instead from complex socio-technical systems in which designers, organizations, operators, and algorithmic processes collectively produce decisions and risks. This shift requires abandoning purely anthropocentric models of accountability in favor of systemic and governance-oriented frameworks capable of capturing distributed agency and shared control.

The analysis of fault reveals that negligence doctrine must evolve beyond individual misjudgment toward institutional and structural responsibility. In algorithmic environments characterized by uncertainty, opacity, and continuous learning, foreseeability cannot be equated with prediction of specific outcomes. Rather, it must be understood as awareness of systemic risks and the obligation to design, monitor, and govern technological systems responsibly. The reasonable person standard must similarly adapt, giving way to context-specific standards that reflect the expertise, authority, and responsibilities of designers, deployers, and institutional decision-makers.

The concept of intent undergoes even deeper transformation. Hybrid cognition fragments intentionality across human goals, machine inferences, and organizational constraints, destabilizing traditional mens rea doctrine. Legal responsibility must therefore recognize degrees and distributions of intentional contribution rather than presuming a unified mental state behind each harmful outcome. Delegated decision authority further intensifies this challenge by separating legal responsibility from practical control. New models of shared and layered responsibility provide a promising path forward by allocating duties and liabilities across the entire socio-technical network that shapes decision outcomes.

Causation, perhaps the most technically strained doctrine, demands a fundamental reconfiguration. Linear cause-and-effect reasoning collapses in the face of probabilistic models, emergent behavior, feedback loops, and multi-actor causal chains. Algorithmic systems produce outcomes through statistical inference and dynamic interaction rather than deterministic sequences. The inadequacy of the but-for test and traditional proximate cause analysis exposes the need for alternative frameworks capable of addressing systemic risk production and distributed influence. Risk-based causation, systemic causation, and governance-based attribution models offer conceptual tools for preserving accountability without oversimplifying causal complexity.

Together, these transformations reveal a common structural conclusion: legal responsibility in the age of human–AI collaboration must shift from event-based blame toward system-based governance. Responsibility must be understood not merely as a backward-looking judgment about individual misconduct but as a forward-looking institutional project concerned

with shaping technological environments, distributing risks, and preventing harm. Law must therefore integrate accountability into the design, deployment, and oversight of AI systems rather than relying exclusively on post-hoc adjudication.

This does not entail abandoning the moral foundations of responsibility. On the contrary, fairness, accountability, and justice remain essential normative commitments. What must change is the conceptual machinery through which those commitments are realized. Legal systems must learn to govern complexity without surrendering accountability, to manage uncertainty without dissolving responsibility, and to accommodate technological innovation without undermining human dignity and social trust.

The future of legal responsibility will thus depend on the capacity of law to reconceive itself as a system of socio-technical governance. Courts, regulators, organizations, and designers will increasingly function as co-authors of responsibility frameworks, shaping how risks are anticipated, distributed, and managed across hybrid decision environments. The task is not merely doctrinal refinement but architectural redesign: the construction of a new legal infrastructure capable of sustaining accountability in a world where human and machine intelligence are permanently intertwined.

If law fails to undertake this transformation, it risks becoming normatively obsolete and practically ineffective. If it succeeds, it may provide one of the most significant contributions to the ethical and institutional governance of technological civilization.

## Ethical Considerations

All procedures performed in this study were under the ethical standards.

## Acknowledgments

Authors thank all participants who participate in this study.

## Conflict of Interest

The authors report no conflict of interest.

## Funding/Financial Support

According to the authors, this article has no financial support.

## References

Ahsen, E., Ayvaci, M., & Mookerjee, R. (2023). When Machines Will Take Over? Algorithms for Human-Machine Collaborative Decision Making in Healthcare. https://doi.org/10.24251/hicss.2023.697

Ardichvili, A. (2022). The Impact of Artificial Intelligence on Expertise Development: Implications for HRD. *Advances in Developing Human Resources*, *24*(2), 78-98. https://doi.org/10.1177/15234223221077304

Arsenijević, B. (2023). Odgovornost Za Štetu Od Veštačke Inteligencije. https://doi.org/10.56461/zr_23.ons.08

Bienefeld, N., Keller, E., & Grote, G. (2023). Human-Ai Teaming in Critical Care: A Comparative Analysis of Data Scientists' and Clinicians' Perspectives on AI Augmentation and Automation (Preprint). https://doi.org/10.2196/preprints.50130

Brand, D. (2022). Responsible Artificial Intelligence in Government: Development of a Legal Framework for South Africa. *JeDEM - EJournal of EDemocracy and Open Government*, *14*(1), 130-150. https://doi.org/10.29379/jedem.v14i1.678

Carter, S. M., Rogers, W., Win, K. T., Frazer, H., Richards, B., & Houssami, N. (2020). The Ethical, Legal and Social Implications of Using Artificial Intelligence Systems in Breast Cancer Care. *The Breast*, *49*, 25-32. https://doi.org/10.1016/j.breast.2019.10.001

Eck, M. V., & Agbeko, F. (2023). Electronic Persons in Contracts. *Obiter*, *44*(4). https://doi.org/10.17159/obiter.v44i4.17507

Gillespie, T. (2022). Building Trust and Responsibility Into Autonomous Human-Machine Teams. *Frontiers in Physics*, *10*. https://doi.org/10.3389/fphy.2022.942245

Hajnal, Z. G. (2021). Artificial Intelligence for Smart Systems Critical Analysis of the Human Centered Approach. *Journal of Computing and Natural Science*, 85-92. https://doi.org/10.53759/181x/jcns202101013

Hasan, A. R. (2022). Artificial Intelligence (AI) in Accounting &Amp;amp; Auditing: A Literature Review. *Open Journal of Business and Management*, *10*(01), 440-465. https://doi.org/10.4236/ojbm.2022.101026

Hill, E. R., Mitchell, C., Brigden, T., & Hall, A. (2023). Ethical and Legal Considerations Influencing Human Involvement in the Implementation of Artificial Intelligence in a Clinical Pathway: A Multi-Stakeholder Perspective. *Frontiers in Digital Health*, *5*. https://doi.org/10.3389/fdgth.2023.1139210

Hryciw, B. N., Fortin, Z., Ghossein, J., & Kyeremanteng, K. (2023). Doctor-Patient Interactions in the Age of AI: Navigating Innovation and Expertise. *Frontiers in Medicine*, *10*. https://doi.org/10.3389/fmed.2023.1241508

Hryciw, B. N., Seely, A., & Kyeremanteng, K. (2023). Guiding Principles and Proposed Classification System for the Responsible Adoption of Artificial Intelligence in Scientific Writing in Medicine. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1283353

Jing-jing, W., Mao, W., & Wenjie, W. (2023). The Ethics of Artificial Intelligence: Sociopolitical and Legal Dimensions. *Isslp*, *2*(2), 27-32. https://doi.org/10.61838/kman.isslp.2.2.6

Kalra, K. (2022). The Future Is Now: Robots as Surgeons the Adoption of Surgical Safety Standards to Robotic Surgery. *International Journal of Law Ethics and Technology*, *2022*(1), 1-15. https://doi.org/10.55574/jymx7781

Lang, B., Nyholm, S., & Blumenthal-Barby, J. (2023). Responsibility Gaps and Black Box Healthcare AI: Shared Responsibilization as a Solution. *Digital Society*, *2*(3). https://doi.org/10.1007/s44206-023-00073-z

Leslie, D., Burr, C., Aitken, M., Cowls, J., Katell, M., & Briggs, M. (2021). Artificial Intelligence, Human Rights, Democracy, and the Rule of Law: A Primer. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3817999

Mohanasundari, S. K., Kalpana, M., Madhusudhan, U., Vasanthkumar, K., Rani, B., Singh, R., Vashishtha, N., & Bhatia, V. (2023). Can Artificial Intelligence Replace the Unique Nursing Role? *Cureus*. https://doi.org/10.7759/cureus.51150

Mühling, Ș. M. K. (2023). Utilizing Artificial Intelligence (AI) for the Identification and Management of Marine Protected Areas (MPAs): A Review. *Journal of Geoscience and Environment Protection*, *11*(09), 118-132. https://doi.org/10.4236/gep.2023.119008

Neves, I. (2023). The Freedom to Conduct a Business as a Driver for AI Governance. *Unio - Eu Law Journal*, *9*(2), 13-56. https://doi.org/10.21814/unio.9.2.5487

Putranti, D., & Anggraeny, K. D. (2022). Inventor's Legal Liability Upon the Invention of Artificial Intelligence in Indonesia. *Varia Justicia*, *18*(1), 71-83. https://doi.org/10.31603/variajusticia.v18i1.6407

Sabt, I. A., & Farooqui, M. O. (2023). Navigating the Convergence of Artificial Intelligence and Space Law: Challenges and Opportunities. *Hightech and Innovation Journal*, *4*(1), 55-64. https://doi.org/10.28991/hij-2023-04-01-04

Santos, F. C. C. (2023). Artificial Intelligence in Automated Detection of Disinformation: A Thematic Analysis. *Journalism and Media*, *4*(2), 679-687. https://doi.org/10.3390/journalmedia4020043

Sio, F. S. d., & Mecacci, G. (2021). Four Responsibility Gaps With Artificial Intelligence: Why They Matter and How to Address Them. *Philosophy & Technology*, *34*(4), 1057-1084. https://doi.org/10.1007/s13347-021-00450-x

Smith, H., & Fotheringham, K. (2020). Artificial Intelligence in Clinical Decision-Making: Rethinking Liability. *Medical Law International*, *20*(2), 131-154. https://doi.org/10.1177/0968533220945766

Sqalli, M. T., Aslonov, B., Gafurov, M., & Nurmatov, S. (2023). Humanizing AI in Medical Training: Ethical Framework for Responsible Design. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1189914

Veitch, E., Dybvik, H., Steinert, M., & Alsos, O. A. (2022). Collaborative Work With Highly Automated Marine Navigation Systems. *Computer Supported Cooperative Work (Cscw)*, *33*(1), 7-38. https://doi.org/10.1007/s10606-022-09450-7

Yang, C. C. (2022). Explainable Artificial Intelligence for Predictive Modeling in Healthcare. *Journal of Healthcare Informatics Research*, *6*(2), 228-239. https://doi.org/10.1007/s41666-022-00114-1

Караджоянни, А. (2021). Research Design for an Integrated Artificial Intelligence Ethical Framework. *The Monitoring of Public Opinion Economic&social Changes*(1). https://doi.org/10.14515/monitoring.2021.1.1911