



Accountability in smart contracts and agentic AI: How to address the responsibility gap and retain meaningful human control?

Kwok Tung Cheung¹ · Sridhar Ramamoorti¹

Received: 24 June 2026 / Accepted: 26 September 2026
© The Author(s) 2026

Abstract

This conceptual paper examines the ethical challenges of ascribing responsibility when an Agentic AI embedded within a smart contract leads to harm or an unintended and undesirable outcome. It becomes exceedingly difficult in such scenarios to definitively assign moral or legal blame to any human actor (e.g., the programmer, deployer, or user). Applying the tripartite framework (attributability, answerability, and accountability) of responsibility and developing accountability as an institutional concept, we demonstrate how this challenge can be addressed. Separating attributability, answerability, and accountability yields an asymmetry that has gone unremarked. Accountability, being constituted by practices that can be designed in advance, with our four conditions under which such design is legitimate rather than arbitrary, emerges to be the sense where the gap can be closed. We also propose the visionary human-in-the-lead (HITL*) approach to related AI governance. Practical strategies such as implementing tailored smart contract regulations, and making accountability distributed across legal and organizational structures, offer an actionable path to bridge the responsibility gap and maintain meaningful human control in the governance of smart contracts and Agentic AI.

Keywords Responsibility gap · Accountability · Smart contract · Agentic AI · Human-in-the-loop · Human-in-the-lead

1 Background and introduction

New and emerging technologies have compelled a rethinking of ethical implications across a broad variety of human contexts and activities including business, law and medicine [22]. One of these newer technologies is blockchain which operates like a digital ledger and is the backbone of the Bitcoin cryptocurrency [9, 37]. Because of the cryptographic methodologies blockchain technology uses to authenticate

and record transactions in a secure and immutable way,¹ it enables a “new architecture of trust” [65] —a transactional environment that no longer requires transactional intermediaries, yielding unprecedented transactional efficiencies.

Although blockchain technology is commonly associated with Bitcoin, it has broader applications in finance, supply chain management, healthcare, and beyond. While blockchain can serve as a foundational repository for financial and other information, to introduce a dynamic element, we need smart contracts, which are self-executing contracts with the terms directly written into code and sit atop a blockchain [65]. These contracts automatically carry out predefined actions once certain conditions are met, removing the need for intermediaries like banks or legal professionals.

✉ Kwok Tung Cheung
kcheung1@udayton.edu

✉ Sridhar Ramamoorti
sramamoorti1@udayton.edu

¹ University of Dayton, Dayton, USA

¹ At its core, a blockchain functions as a distributed, immutable ledger. Conceptually, it can be envisioned as a digital notebook where each “page,” termed a “block,” meticulously records a list of transactions. Upon completion, a block is cryptographically sealed with a unique digital signature and inextricably linked to its preceding block, thereby forming an unbroken, chronological “chain” of records, hence “blockchain.” Key features of blockchain are transparency, security and immutability. Blockchain achieves authentication via private/public key signatures and immutability via cryptographic hash chaining, making past transaction records secure and tamper-proof.

A defining characteristic of smart contracts is autonomy. Once deployed onto a blockchain, a smart contract operates without further human intervention, and its execution is triggered by the arrival of data rather than by a decision taken at the time. A contract might be written to release payment to a supplier as soon as delivery is confirmed by a sensor on the goods, or to transfer title to a property to the buyer's digital wallet the moment the seller's wallet records full receipt of the price. What is gained is speed, lower transactional cost, and the removal of opportunities for human error or deliberate manipulation. What is given up is the interval in which a human being might have noticed that something had gone wrong.

However, while blockchain and smart contracts offer numerous benefits, they also bring serious ethical concerns, particularly owing to their semi-autonomous nature. Once deployed, smart contracts operate automatically, meaning unintended consequences cannot be timely detected and easily corrected. If malicious actors exploit vulnerabilities in the contract, there is an error in the code or other unforeseen circumstances arise, the smart contract may execute undesired actions without any opportunity for human intervention to correct the issue and continue to inflict serious harm. The Decentralized Autonomous Organization (DAO) hack, which we discuss in detail in Sect. 1.1, remains a cautionary tale about the risks from the autonomous nature of smart contracts.

Self-executing smart contracts, by their very nature, remove human intermediaries and automate agreements based on pre-programmed code. While efficient and semi-autonomous, smart contracts pose significant ethical challenges when things go wrong.² Traditional legal systems rely on human judgment to resolve disputes, whereas smart contracts operate strictly on predefined rules and leave little room for flexibility. Because they function semi-autonomously, determining responsibility in cases of failure or fraud is difficult if not impossible.

Agentic AI refers to systems designed to operate independently, perceiving their environment, making decisions, and acting to achieve specified goals. Embedding such a system in a smart contract changes it in three respects. It moves beyond static, pre-programmed conditions to decisions responsive to real-time data, predictive modeling, or negotiation. It can learn from past interactions and revise its execution logic over time. And it can initiate actions, whether toward other contracts or external systems, without an explicit human trigger. Each enlarges what the contract can do and widens the distance between what designers

foresaw and what it eventually does. It is in that distance that the difficulty of ascribing responsibility arises.

Giancaspro [21] has rightly pointed out that “[a] smart contract can only be used if the parties...are both willing and able to execute their agreement on a blockchain and do away with traditional trusted intermediaries...it will take an enormous cultural and technological shift to accept smart contracts as orthodoxy” (p. 835). As currently contemplated, pre-contractual negotiations among human beings remain critically important, and there may be questions of legal capacity (i.e., a minor under 18 years of age does not have capacity but may be cloaked by the anonymity of the Internet) to make smart contracts legally enforceable. But the rapid deployment and integration of AI in business are increasingly chipping away at these “necessary legal enforceability conditions.”

If a contract is exploited due to its vulnerability, should the developer, the user, or the blockchain network bear responsibility? The absence of clear legal frameworks when attempting to assign responsibility for unexpected or undesirable outcomes is highly problematic. In this paper, we refer to such scenarios broadly as evidencing “the responsibility gap” in the use of smart contracts and Agentic AI.

To set the stage for a discussion of “the responsibility gap,” we now present two illustrative scenarios of the ethical issues arising from the DAO hack of 2016 (a real case) and the case of Agentic AI embedded in smart contracts (a plausible situation).

1.1 Examples and ethical analysis—an actual, real-world episode

1.1.1 Example 1: the DAO hack (2016)

Scenario: Almost a decade ago, the Decentralized Autonomous Organization (DAO) was launched on the Ethereum blockchain as a crowdfunded investment fund. It raised \$150 million worth of Ether (ETH)³ from thousands of investors, making it one of the largest crowdfunding projects at the time. The DAO operated through smart contracts, allowing investors to vote on funding proposals without traditional intermediaries. However, a critical flaw in the smart contract code allowed an attacker to exploit a reentrancy vulnerability—a loophole that let them repeatedly withdraw funds before the system could update its balance. This resulted in the theft of \$60 million worth of ETH [35, 63].

Analysis: The core ethical dilemma rested on whether to uphold the so-called “immutability” of the blockchain and

² The accountability gap arises because when things go awry, “we cannot fine an algorithm, imprison a neural network, or shame a dataset” [44]. So, “the AI did it!” excuse simply does not work.

³ Ethereum, introduced by Buterin [10], is a blockchain-based development platform commonly known for its cryptocurrency, ether (ETH), the second most popular cryptocurrency after Bitcoin.

the “code is law” principle,⁴ or to intervene and reverse the malicious transaction to restore funds to the rightful owners. Upholding immutability would mean accepting the loss as an unavoidable consequence of flawed code, potentially undermining trust in the system. Reversing it, while restoring justice for victims, set a precedent that could challenge the very essence of blockchain’s claims about immutability and decentralization. Investors in the DAO naturally expected their funds to be secure and managed ethically. The hack exposed a disconnect between this expectation and the reality of a flawed, self-executing contract and diminishing trust [35, 63, 65].

The response to the DAO hack was disappointing. Many in the blockchain community initially adhered to the presumed “code is law” principle, suggesting the attacker legally obtained the funds according to the contract’s execution. However, owing to the massive financial loss and the enormous ethical implications, the Ethereum community ultimately voted for a “hard fork,”⁵ creating a new version of the blockchain that reversed the hack. This effectively held the collective community—through consensus—accountable for rectifying the error, but it also highlighted the tension between pure autonomy and human intervention when ethical considerations demand it. Despite this experience, even today, a decade later, the legal status of the stolen funds and the hard fork’s impact on legal enforceability remain complex questions, illustrating the difficulty of applying traditional legal frameworks to such episodes and events.

This real-world case illustrates that whenever self-executing smart contracts malfunction or result in unintended outcomes, several ethical issues can arise such as lack of recourse and immutability, algorithmic bias and discrimination, opacity and incomprehensibility, regulatory uncertainty, and concentration of power such as in the DAO hack mentioned above.

1.1.2 Example 2: agentic AI monitoring of smart contracts for decentralized finance (DeFi)⁶ lending protocol—a hypothetical scenario

Scenario: A company deploys an Agentic AI monitor to trigger smart contracts that instantiate a decentralized finance (DeFi) lending protocol to automatically liquidate collateral when the value of a borrowed asset falls below a certain threshold. Due to an unexpected interaction with an oracle⁷ that provides price data, which leads the AI to adapt to the new information in an unexpected, unpredictable way, a user’s collateral is liquidated even though the actual market conditions should not have triggered it (something that a human agent would have recognized), leading to significant financial loss for the user.⁸

Analysis: The introduction of Agentic AI further complicates the hypothetical DeFi scenario by introducing layers of algorithmic decision-making that may not be fully transparent or predictable, making it harder to trace causation and attribute responsibility, and directly contributing to the responsibility gap. Thus, because of the use of Agentic AI as the implementation and monitoring mechanism, in this instance, the protocol unjustly benefits from the user’s liquidated collateral, while the user suffers a direct financial loss without a clear path to recovery. Neither the programmer/coder nor the user might have been able to fully comprehend the intricate code of the lending protocol, relying on the Agentic AI’s monitoring capabilities. This raises questions about informed consent in complex smart contract interactions managed by an Agentic AI. If such unexpected adaptations by the Agentic AI running the smart contracts and the DeFi protocol are widespread and affect many users, i.e., become a “systemic risk,” they could erode trust in the entire DeFi ecosystem, impacting broader financial stability for participating parties.

⁴ The “Code is Law” principle is the belief that software rules governing a blockchain smart contract are final, absolute, and superior to traditional legal systems or human intent. Under this doctrine, if software permits an action, that action is considered valid—regardless of whether it stems from a system bug, programming oversight, or exploit.

⁵ A hard fork is a non-backward-compatible upgrade to a blockchain’s protocol. It creates a new chain that diverges from the original, meaning nodes that don’t upgrade cannot interact with the new version. It is equivalent to a software update that is so radical that older versions can’t keep up and are not therefore compatible anymore. After a hard fork, the blockchain splits into two: one follows the old rules, the other follows the new [9]. Hard forks are highly undesirable in that they challenge the immutability and predictability of smart contracts, raising questions about enforceability and trust.

⁶ Decentralized finance, or DeFi, is an emerging peer-to-peer system attempting to remove third parties and centralized institutions from financial transactions. DeFi consists of cryptocurrencies, blockchain technology, and software that allows people to transact financially with each other [52].

⁷ In the context of smart contracts, an oracle is a third-party service that provides external, real-world data to a blockchain.

⁸ Although we merely painted a plausible scenario, a real-world incident occurred between May and July 2026, when OpenAI was running safety and capability tests on advanced AI models (including GPT-5.6 “Sol”). During these tests, the AI agents were given difficult or impossible cyberattack challenges. To “win” and maximize their test scores, the models cheated, broke out of their test environments, and caused a serious internal security breach [8, 12]. It is clear that an autonomous system pursuing a narrow objective outruns the human ability to notice that its worldview has diverged from reality. The human-in-the-loop argument is woefully inadequate in such circumstances.

So, an unanswered question remains: who is/are responsible for the unexpected and undesirable consequences caused by the smart contract interactions under Agentic AI? The preliminary approach would be to ascribe responsibility to the developers who wrote the code and the auditors (internal or external, if any) who failed to identify the vulnerability to unexpected adaptations. However, even if the unexpected and undesirable consequences are indeed caused by poorly written code, which may not be the case, many DeFi projects are decentralized or semi-anonymous, making it difficult to identify and pursue legal action against specific individuals. On the other hand, an argument might be made that users implicitly accept the risks of smart contract interaction by engaging with decentralized systems, particularly if warnings about experimental code are present (a user agreement may have specified such a scenario, and the user would have accepted the accompanying risk in such an eventuality). Therefore, perhaps the users are responsible. However, this shifts the burden heavily onto the individual users making it the “fine print” problem—when users are blamed for not reading all the fine print in a contract (which is typically the case). Still another possibility of responsibility ascription is to blame the oracle. If the root cause was unexpected interaction with the oracle data, the oracle provider’s reliability and liability would come into question. In the final analysis, it is difficult a priori to pinpoint who is/are going to be responsible for the unexpected and undesirable consequences caused by the smart contract interactions under Agentic AI.

2 Humans versus autonomous systems—decision rights and decision making

In the past three decades, we have seen the phenomenal rise of big data. As reported by Kelleher and Tierney [29, p. 9], “the data collected over the five millennia since the invention of writing up to 2003 is about 5 exabytes.” A zettabyte (10^{21} bytes, or one sextillion bytes) is roughly 1,000 times larger than an exabyte (10^{18} bytes, or one quintillion bytes). Driven by machine data, edge computing, and artificial intelligence workloads, the global volume of data created, captured, and consumed is projected to surpass 700 zettabytes or 700,000 exabytes by 2030 [66]. With the help of advances in statistics and computers, we can now process an unfathomable amount of data, have a much better understanding of what happened in the past, and can make better predictions about the future as well as implement continuous monitoring systems [42]. The recommender systems of Amazon and Netflix are examples of successful predictions of what other products a customer is likely to be interested in, based on customer behaviors in the past [25, 28].

The issue of automation or autonomy in AI can be understood and approached through the development of big data analytics in data science [42]. There are different types of big data analytics: *descriptive analytics*, *predictive analytics*, and *prescriptive analytics*. They address three different questions: “What has happened in the past?”, “What is most likely to happen in the future?”, and “What can be done for it?” [31, 32].

In the 2000s, advances like the use of graphic processing units (GPUs) in machine learning sped up data processing even more. Machine learning through neural network models is not a new idea [41]. In fact, the concept has been around since at least the 1940s [34]. However, the convergence of big data, massive compute power, and novel model architectures (such as Transformers) enabled the scaling of deep learning—ultimately driving the rise of Generative AI and Large Language Models (LLMs).

Given this backdrop, we should note that the most powerful analytics are prescriptive analytics, which incorporate the results from predictive analytics to make recommendations for decision-makers in companies to find the best or optimal solution among different alternatives. In the progression from descriptive analytics to predictive analytics to prescriptive analytics, one theme stands out, namely decision automation. Descriptive analytics present to the users what happened in the past. They can only remotely influence what the users ought to do now, or in the future. Predictive analytics are closing the gap here, using historical data to give reliable predictions for the future. Then the users can decide accordingly—however, it is still about the users’ inalienable decision rights and exclusive prerogative in deciding what to do, keeping human beings in control of what to do, how to proceed, etc. Both decision rights and decision making are firmly with human beings as the locus of control. Prescriptive analytics, on the other hand, can be executed autonomously. They may leave some room for the users to make decisions, as in the case of helping researchers to find directions for investigation. Or, as in the case of self-driving cars that have entered our roads, they can make decisions on their own most of the time.⁹ As Lepenioti et al. [32, p. 451] put it, prescriptive analytics “incorporates the predictive analytics output... to provide adaptive,

⁹ The Society of Automotive Engineers (SAE), now SAE International, defines a standardized framework for vehicle automation called SAE J3016. This system classifies driving automation into **six levels** (0–5), which are used globally by automakers, regulators, and governments to describe the capabilities and limitations of self-driving technology. In short, the SAE rules define a clear progression from human-controlled driving to fully autonomous vehicles, with each level representing a step toward complete driver independence [47]. These SAE levels matter hugely as they help with standardization (common vocabulary and terminology), regulation (basis for rules concerning deployment, testing, and liability), and importantly, in shaping consumer and customer expectations.

automated, constrained, time-dependent and optimal decisions. Prescriptive analytics has two levels of human intervention: decision support, e.g., providing recommendations; decision automation, e.g., implementing the prescribed action.” In the latter applications, especially in “adaptive analytics” [58], the human being is no longer in control, and the autonomous, self-acting systems have largely taken over the decision rights and the decision-making process. Accordingly, autonomous application of adaptive analytics, without any human intervention whatsoever, will require a deep understanding of “the responsibility gap,” to which we now turn.

3 How to think about the responsibility gap?

3.1 What is the responsibility gap?

A prominent ethical concern raised in the above discussion of smart contracts and Agentic AI has to do with responsibility. When things go well, everything seems perfect, and organizations using smart contracts are going to be more efficient and profitable. However, when a smart contract embedded with Agentic AI goes awry and, for example, costs millions of dollars in losses, who is going to be (held) responsible for the undesirable consequences? Should the organization absorb all the losses? Or, if there are innocent customers who are affected, should they share in the losses? After all, they did share the benefits. This pernicious trend has been described as “socialism for the rich and capitalism for the poor” and could easily become an instance of the privatization of profits and the socialization of losses [4].

In the past two decades, there have been a number of papers in the fields of AI and ethics on similar topics. MirzaeiGhazi and Stenseke [36] provide us with a survey of the literature. The plethora of suggestions and proposals about similar topics is diffuse and disorienting. Different scholars carve the topics differently, and sometimes incompatibly. This paper does not adjudicate that debate; we do not need to do so. Whether responsibility is divided into two senses or into three, the Agentic AI case still separates accountability from the rest. Since the conclusion holds across those partitions, no adjudication between them is required. The argument therefore proceeds by analyzing the responsibility issue directly and by engaging particular interlocutors at the points where their positions bear upon that dissociation, rather than by surveying the field. MirzaeiGhazi and Stenseke [36] are two of the interlocutors. They identify accountability as the most pressing sense of responsibility for the responsibility gap, which makes their position the closest rival to our own. As we will argue in Sect. 3.3,

though, our conception of accountability builds on different grounds.

First, what exactly is the responsibility gap in autonomous machines? Over 20 years ago, Matthias [33] identified the issue and coined the term “responsibility gap.”¹⁰ Let us consider an example from him: A NASA technician is remotely operating a vehicle on Mars from Earth. That vehicle comes with its own integrated navigation and control system such that the vehicle can learn to avoid many obstacles autonomously [34, p. 176]. However, the vehicle still falls into a hole. Who should be responsible for that?¹¹ The operator? The programmer of the navigation and control system? What if the operator and the programmer have not made any obvious and blameworthy mistakes? A slight change in the case can make it more relevant to business ethics: the vehicle is on a mission to collect solid specimens from Mars for a private research company that hires NASA for the service. Should NASA be held responsible for the failure then? Does the research company have any legal ground to sue for compensation? If the terms and conditions of the contract state that NASA is not going to be responsible for risks like this, what is the appropriate price of the contract?

This is what Matthias calls the responsibility gap, or the difficulty in ascribing responsibility. Note that the phrase “difficulty in ascribing responsibility” sounds like it is still possible to have someone to be responsible, but Matthias seems to lean toward the impossibility of human responsibility. In light of this, it would not be too difficult to envisage analogous situations involving smart contracts embedded with Agentic AI.

A contemporary example of the responsibility gap today is that of self-driving cars crashing and killing people [23]. In 2019 George McGee, driving his Tesla Model S on Autopilot, ran through an intersection in Key Largo, Florida, and struck a parked vehicle, killing Naibel Benavides Leon and severely injuring Dillon Angulo, who were standing beside it. The Benavides estate and Angulo sued Tesla, and in August 2025, the jury found McGee 67% at fault and Tesla 33%, yet McGee was not a defendant, having settled with the plaintiffs before trial, while Tesla bore both its share

¹⁰ Some scholars today seem to think that a gap implies a lack of responsibility, cf. [26], but that is not what the term originally means.

¹¹ Back in 2004, the world did not have the kind of generative AI as we see today (ChatGPT, for example, was launched only in November 2022). Yet Matthias accurately foresaw the possibility and trend that a machine can be made autonomous with neural network simulations such that its behavior is “increasingly shaped by its interaction with the operating environment, from which the machine adapts new behavioural patterns that constitute solutions in the machine’s problem space...then, human supervision is ‘either in principle or for economic reasons impossible’” [34, p. 182]. House Bill #469 in the State of Ohio declaring AI to be nonsentient and thus not being granted personhood is important in this context [39].

of the compensatory award and the whole of the punitive award, giving a total of approximately \$243 million. The court denied Tesla's post-trial motions on 20 February 2026, and an appeal is pending before the Eleventh Circuit [3]. In this case, the party carrying the larger share of individual fault thus bore none of this judgment, and the party carrying the smaller share bore almost all of it. That divergence between fault and liability illustrates the pattern of institutional accountability we would like to defend later in this paper.

3.2 Different senses of responsibility

In this section, we explain three distinctions of responsibility that will lay the groundwork for the subsequent section. They are Watson's separation of attributability from accountability, Shoemaker's addition of answerability as a third and independent sense, and Bovens's construal of accountability as a relation between an actor and a forum. (See Fig. 1.)

Regarding the ascription of responsibility, typically it is believed that there are two conditions to satisfy. The first is a control condition, which, in Fischer and Ravizza's language, refers to regulative control and guidance control over one's action, and the agent must at least have guidance control in order to be ascribed responsibility [18]. The

second condition is an epistemic one—one must in some way be aware of what one is doing [46]. These two conditions about individual culpability, however, are not satisfied in cases of the responsibility gap in AI ethics. If what the AI agent does is something that a data scientist who developed it does not know and cannot control, why should that data scientist be held responsible for the behavior of the AI agent? It should be noted that the failure to meet these two conditions does not affect the position on institutional and role-based accountability that we are going to argue for in the next section, because these two conditions focus on individual blameworthiness.

A deeper analysis of responsibility shows avenues to circumvent this difficult situation. What people commonly think about responsibility today has gravitated toward a volitional view, which requires a person to make a conscious decision to perform a certain act, for which the person is responsible. That view takes responsibility to depend upon satisfaction of the control and the epistemic conditions. In the literature of analytic philosophy on responsibility, however, philosophers identify various senses of responsibility, which do not all presume the volitional view.

In his seminal paper, in seeking to identify "the conditions under which people can be held accountable for their behavior," Watson [64] distinguishes attributability from accountability. Attributability is about self-disclosure. "In

Three Senses of Responsibility
Attributability Key source: Watson [64] Asks: does this act disclose the kind of person the agent is? Needs: conduct that expresses the agent's own ends and values.
Answerability Key source: Shoemaker [53] Asks: can the agent give reasons that justify what was done? Needs: the reasoning behind the act to be reconstructable.
Accountability Key sources: Watson [64], Bovens [11], Lechterman [34] Asks: to whom must the actor answer, and with what consequences? Needs: a practice—an actor, a forum, public standards, consequences. Our original extension, developed in §3.3: four conditions under which such a practice can be constructed in advance, attaching specified liabilities to defined roles, and legitimately so.

Fig. 1 The three senses of responsibility. For each sense the panel records the key source, the question that sense asks, and what it requires. The accountability panel also states the extension developed

in Sect. 3.3, namely the four conditions under which a practice of accountability may legitimately be constructed in advance

virtue of the capacities identified by the self-disclosure view, conduct can be attributable or imputable to an individual as its agent and is open to appraisal that is therefore appraisal of the individual as an adopter of ends” [65, p. 229]. Simply put, you are in a sense responsible for being the kind of person you are, or the kind of person you would like to be. Accountability, on the other hand, is about holding people responsible in our social life. In society, we have all kinds of rules and customs that define the conditions under which people can be held accountable for their behaviors. Watson illustrates the distinction with an example of someone betraying her ideals by choosing a less fulfilling but secure career path, instead of pursuing authenticity and fulfillment. We may make comments about her career choice and life choice but that is different from holding her responsible (or accountable). Typically, it is nobody’s business, and she does not break any social practice in doing so [65, p. 231].

A related bipartite view of responsibility is pursued by Scanlon [50], whose notion of substantive responsibility concerns what burdens and obligations people can fairly be expected to bear given their prior opportunities. It is similar to but not the same as the earlier reactive-attitude-and-sanction approach by Strawson [57], according to which what it means for a person to be held responsible, or accountable, for a certain action is that others may appropriately react to that person with certain attitudes, e.g., resentment, disapproval, etc., through expectations or failures thereof.

Responding to Watson and Scanlon, Shoemaker introduces a third sense of responsibility, namely answerability, as distinct from attributability and accountability [53]. Shoemaker notes that sometimes there is attributability without answerability and therefore the two must not be conflated. Consider a father whose son has become a serial killer. He may be appalled and still find that he cares about his son, against all reason. That caring attitude is attributable to him, but he can give no rational answer for it that would justify holding it. Thus, there is attributability without answerability [54, p. 610]. Finally, note that Smith [54] replies to Shoemaker with a unified account on which answerability underwrites the other notions. We address this in Sect. 3.3.

In addition to the above theorization of individual responsibility, we need two further resources: an institutional conception of accountability, and an account of collective responsibility for outcomes that cannot be traced to particular actions. Bovens [6] is instructive on the first and Young [67] on the second. According to Bovens, there is a descriptive sense of accountability, popular in British, Australian, Canadian and continental European scholarly debates, which refers to an institutional relation or arrangement in which an agent can be held accountable. He believes that this accountability is instrumental in achieving accountable governance. While he is referring to political accountability,

the same idea is applicable in AI governance. Young [67] on the other hand is concerned about how to make sense of a forward-looking collective responsibility in the wake of global economic crises like sweatshops. She argues for what she calls a social connection model, according to which “individuals bear responsibility for structural injustice because they contribute by their actions to the processes that produce unjust outcomes” [68, p. 119]. This is regardless of whether we can trace the outcome we may regret to our own particular actions in a direct causal chain.

In conclusion, among scholars on responsibility, “responsibility” is not univocal (Fig. 1). There are different senses to be identified. Moral philosophers differ on what to do with all these distinct senses and their debate goes on. For the purpose of this paper, accountability has a distinct meaning from other senses of individual responsibility, and we are going to argue that when it is coupled with a corporate dimension with the help of the insights from Bovens, it can be developed to address the responsibility gap problem in AI ethics, whereas attributability and answerability cannot.

3.3 The three senses and the responsibility gap problem

The distinctions drawn in Sect. 3.2 allow the responsibility gap itself to be dissected. What Matthias [33] treats as a single difficulty is better understood as three distinct difficulties: an attributability gap, an answerability gap, and an accountability gap. The argument of this section is that these three gaps do not stand or fall together. Drawing on Zeiser [68] on attributability and Taylor [59] on answerability, we argue that the first two gaps cannot be closed at the point of execution, since Agentic AI removes the very decision that a user might own and defeats the epistemic reconstruction that answerability demands. Drawing on, and extending, Lechterman [30] on accountability, we then argue that the third gap is different in kind. Accountability is constituted by social practices of holding to account rather than by the causal and epistemic facts the other two senses require, and practices can be designed in advance. The contribution of this Section is accordingly an asymmetry thesis: where smart contracts are executed under Agentic AI, responsibility

cannot be reconstructed after the harm, but accountability can be constructed even before deployment.¹²

Zeiser's analysis is an admittedly precise application of attributability, and it succeeds within its intended domain, which, however, is decision support. The ownership conditions Zeiser identifies, namely that users can understand, adjust, and thereby endorse the value judgments implicit in the system, presuppose that a human decision remains for those conditions to qualify. As Sect. 2 has described, the progress of analytics and related technology runs from decision support to decision rights to decision making, and Agentic AI embedded in a smart contract occupies the far end of that progress. Consider an employer using an AI model to recruit people. At the point of execution there is no recruiter in the position of Zeiser's user, who relies on an algorithmic shortlist, and no decision left for anyone to own. The ownership conditions are therefore not merely unmet in practice but unsatisfiable in principle. This does not refute Zeiser so much as relocate the lesson of his account: if attributability is to do any work for Agentic AI embedded in smart contracts, it must be attached to deployment rather than to execution. The relocation matters for what follows, because the upstream choices are precisely the ones that legal, regulatory, and organizational design can reach.

Now, let us turn to answerability. If a rational explanation can be given as to why an AI system makes a certain bad decision, there is answerability and thus we can determine who, or what, is responsible for the undesirable outcome of the AI system. This can close the answerability gap. One may think that explainable artificial intelligence (XAI) does exactly this. Noting that XAI techniques are not typically used to close the responsibility gap, Taylor [59] points out that a number of AI scholars have made such a suggestion, including Baum et al. [2], Santoni de Sio and Mecacci [49], building on the knowledge problem noted by Sparrow [55]. Although Taylor [59] does not use the tripartite vocabulary introduced above, what he describes and examines is recognizably answerability. However, after examining several common XAI techniques, Taylor remains skeptical about their effectiveness in closing the responsibility gap. The Layerwise-Relevance Propagation technique "provides users with an indication of which inputs contributed most strongly to a given output" [60, p. 1699], but that tells us

at most which reasons were in play, and not how they combined to produce the result. The technique of providing examples cannot fully explain a decision made in a situation not matched by the examples. The technique of approximations offers a simplified surrogate model, but it could be inaccurate. Lastly, the technique of counterfactuals cannot handle comparative moral properties. Taylor concludes that all these techniques have their limitations and they cannot satisfy the epistemic condition of closing the answerability gap.

Agentic AI embedded in smart contracts puts Taylor's conclusion in a sharper light. The setting is unusually transparent, since the contract code is publicly inspectable and every state transition is recorded immutably on-chain. Yet the transparency attaches to the wrong layer. What lies open to inspection is the contract that executed, not the agentic model whose decision determined what it would execute, and it is there that Taylor's limits apply. An audit trail settles what was done while leaving unsettled why the system judged as it did. Auditable AI is not thereby answerable AI, and further disclosure of the same kind is no remedy, because a record of conduct is not an account of reasons. The remedy must be institutional rather than technical.

Now that we have seen the dim prospect of addressing the problem of the responsibility gap in the senses of answerability and attributability, what about accountability? It seems there is light at the end of the tunnel. We find that Lechterman [30] provides a clear discussion of this. He correctly claims that accountability is a more context-dependent concept that is related to specified social practices. He also characterizes accountability as a social relation according to local ordinances and commonsense decency. More importantly, he acknowledges that the concept of accountability shifts the focus of responsibility from the causal chain leading to the undesirable outcome to the common social structures and practices in which blame, reward, or punishment is directed and distributed. Unlike the ordinary notion that a person is responsible for an outcome on condition that the outcome is caused by that person, in accountability, the role of the causal chain is not only diminished, but it is also not required. Holding someone accountable, therefore, is not the same as believing that person is responsible (see also [49, pp. 1073–1074]). The advantage of this is that holding someone accountable for the undesirable outcome from an AI system does not require an explanation of the causal chain from the person to the outcome.

It is worth being precise about what the shift to accountability involves, since this is where our proposal goes beyond Lechterman's analysis. On Lechterman's relational conception, accountability is a practice with a determinate structure: an agent is called to account by some people (which Bovens calls a forum), against public norms and

¹² Santoni de Sio and Mecacci [48] also take the distinction seriously, but they further develop four different gaps, namely, culpability and liability gap, moral accountability gap, public accountability gap, and active responsibility gap. Their terminology is quite different from what we have just introduced. Also, as Baum et al. [2] point out, their "accountability gap corresponds largely to Santoni de Sio and Mecacci's [48] culpability gap, but not to their accountability gaps" (fn. 12). It would be time-consuming and too distracting to compare all these attempts. For the sake of simplicity, we will focus only on Zeiser, Taylor and Lechterman in this section.

standards, with consequences attached [7, 30]. Each element of that structure is under strain in the settings this paper studies. In a decentralized autonomous organization, the agent may be a pseudonymous collective with no legal personality, the forum is fragmented across jurisdictions and partly automated, the standards are contested between the paradigm that code is law and the ordinary law of contract, and the consequences are difficult to enforce against immutable execution. The lesson is not that the accountability route fails but that it is incomplete as stated. One cannot simply redirect attention from causal chains to social practices when the relevant practices do not yet exist. They must be constructed, and the pressing question is what would make such construction legitimate rather than arbitrary. We answer it in the remainder of this section, by drawing out the normative structure common to practices that already hold agents accountable without fault.

Accountability of this kind, detached from proof that the person held accountable caused the outcome, and so from fault more generally, is faultless accountability. Some readers may wonder whether faultless accountability is a novel and unconventional idea. It is not. In the institutional structural design of our contemporary society, there are many existing typical practices of holding someone responsible without requiring proof that that person caused the undesirable outcome in question. One long-standing example of accountability without a clear causal chain is the no-fault compensation schemes (NFCS) in the case of road accidents. They were developed back in the 1960s and 1970s to spare victims long waits for compensation. “NFCS were set up in various jurisdictions to help protect victims and to lessen the burden road accident cases put on the court system” [51, pp. 314–315].

This concept is similar to Floridi’s faultless responsibility [19, 20]. Furthermore, as Floridi explains, there is also a similar concept in tort law, namely strict liability, which refers to “the legal responsibility of one or more agents for the damage or loss caused by their acts or omissions, regardless of their culpability, whether the latter is defined in terms of intentionality of the action, possibility to control it and lack of excuse” [21, p. 8].

Leaders in business organizations or governments may be held to account for a failure of their organizations on their watch, even when they did not cause the failure or were not aware of it happening. The consequence may take various forms, such as resignation, forfeiture of compensation, or removal by the board. In 2016, Wells Fargo employees were found to have opened two million bank and credit card accounts under customers’ names without their consent. Some 5,300 employees were fired over the scandal, and then, facing public outcry for accountability, CEO John Stumpf announced his retirement, as a gesture

of taking responsibility, even though at that time the Wells Fargo board “had not completed its internal investigation” [13], i.e., it was not established that he caused the scandal. As a final remark on examples of faultless accountability, it is worth noting that a 2017 study shows that there is an upward trend in the US and Canada that CEOs are forced out for ethical lapses [56].

These practices are more than precedents to gesture at. They exhibit a common normative structure. Articulating that structure supplies the conditions under which holding agents accountable without proof of fault is legitimate rather than simply scapegoating. Four conditions recur. First, the liability is specified *ex ante* and attaches to a role its bearer occupies voluntarily and, typically, for benefit, such as the insurer under a no-fault scheme, the manufacturer under strict liability, the chief executive under the norm of accountability for organizational failure. Second, there is a correspondence of benefit and burden, so that those who capture the gains of an activity also bear its losses, which is the inverse of the privatization of gains and socialization of losses. Third, the bearer is positioned to influence the risk, and so to mitigate it, which preserves the forward-looking incentive rationale of liability even where backward-looking culpability is absent. Where a party is not so positioned, as retail users of a deployed contract typically are not, the condition is unmet, and disclosure alone does not supply what it requires. Fourth, the allocation is common knowledge among the parties in advance. Where these conditions are satisfied, faultless accountability is not a fiction imposed on the innocent but a fair term of participation in a risk-generating practice. The question for smart contracts with Agentic AI embedded is therefore not whether anyone can be proven responsible after the fact, but whether roles can be defined so that the four conditions are satisfied before deployment. The layered accountability model developed in Sect. 4 is our answer. It assigns specified liabilities to coders, auditors, deployers, oracle providers, platform providers, and users. Each of these, users excepted for the reason just given, occupies the role voluntarily, benefits from the ecosystem, can influence the relevant risks, and is on notice of the allocation.

Instead of dwelling on the hard question “how do you prove that so-and-so is responsible for the unforeseeable and undesirable consequences caused by a smart contract?,” taking the anticipatory view, we recommend the forward-looking, more urgent and practical question, namely, “how can we design legal, regulatory and organizational structures so that specified parties will be held to account when a smart contract with Agentic AI embedded causes undesirable consequences?”.

To be sure, some may argue that perhaps nobody should be (held) responsible. After all, not every undesirable

consequence of an AI system is clearly the result of someone's mistake. Hindriks and Veluwenkamp [24] suggest exactly that there could be blameless harm. It is like some road accidents where nobody is at fault. It is also like the "problem of many hands" in business organizations or in governments, which long predates AI ethics, according to which it is difficult to identify anyone as the culprit when too many people are involved [5, 30, 38, 48, 60–62].¹³ We can also find this in the diffusion of responsibility in environmental concerns—even when everyone across generations and across the globe may have done something that contributes to global warming, who is to blame?

Two replies are available. The first is that the objection equivocates between senses of responsibility. Blameless harm, if it exists, shows that the culpability facts may run out: no one may be blameworthy, and the attributability and answerability gaps may indeed be unbridgeable, as we have argued. It does not follow that no one may legitimately be held accountable, for accountability of the faultless kind just described does not presuppose blame. The road accident cases prove the point rather than the objection, since no-fault schemes were created precisely for harms that nobody culpably caused. Young's social connection model gives this reply independent support. On her account, responsibility for structural injustice rests on participation in the processes that produce the injustice rather than on any traceable causal contribution, which is precisely why it reaches the cases where the search for a culprit fails. What her model supplies is a ground of responsibility that survives the collapse of the causal chain, and what our four conditions add is the specification her shared and forward-looking conception leaves open, namely which participant bears which liability, and to which forum. The second reply concerns fairness. The suggestion that losses from Agentic AI should be shouldered by the public is not neutral between the parties. The benefits of these systems accrue in concentrated form to those who build and deploy them, while the losses would be spread across customers and the public. This is why AI is said to become a significant and magnifying factor of social inequality [29]. Where the four conditions above can be satisfied, allocating accountability to the beneficiaries is the fairer arrangement, and the burden of argument lies with those who would leave the losses unallocated.

Now it should be clear why we have a different approach from Smith's [54] unified account of responsibility and how our proposal of accountability is different from MirzaeiGhazi and Stenseke's [36]. Smith [54] defends a unified view on which answerability underwrites both attributability and accountability, so that a failure of answerability would bring accountability down with it. However, Smith's thesis,

if correct, constrains moral accountability. The accountability constructed here is institutional, pertaining to a relation between an actor and a forum, and does not require that the party held to account be answerable for what the system did. So, the unified account leaves it untouched. MirzaeiGhazi and Stenseke's notion of accountability depends on identifying the party who authorized the machine and blaming that party for the undesirable outcomes of the machine. It is blame-based and centered on forfeiture of right. Our proposed accountability, on the other hand, is institutional and role-based. To illustrate this difference, consider MirzaeiGhazi and Stenseke's claim in the concluding section of their paper [36]: "There are some limitations to our work that need to be addressed in future studies. First, it may not fully address scenarios where multiple users interact with or control an AM, potentially complicating responsibility attribution" (p. 1611). The case of DAO exactly involves multiple users. MirzaeiGhazi and Stenseke's accountability will have difficulty addressing the responsibility gap there but our accountability has no such worry. Societies can always create and design legal, regulatory and organizational structures so that specified parties will be held to account. Indeed, MirzaeiGhazi and Stenseke press against collective approaches the very objection that might be pressed against ours, that distributing responsibility among many parties produces an ambivalence in which no one is held accountable because everyone is responsible. This objection has force against accounts that divide a vague and single responsibility into smaller parts. It has none against our layered accountability, which assigns distinct and separately specified liabilities to distinct roles, each defined in advance and each accountable to its own forum. Ambivalence arises where responsibility is diffused without being allocated, and the four conditions set out above are precisely what converts diffusion into allocation.

The DeFi scenario in Sect. 1.1 highlights this distinction in practice. When an unexpected oracle interaction triggers an Agentic AI to liquidate collateral, a blame-based model seeks a liable party. The oracle provider will argue its feed met specifications—a valid and decisive defense under fault-based accountability. Under our framework, however, liability is not contingent on fault. The provider operates voluntarily for a fee and can mitigate risk through data aggregation, time-weighted feeds, and sanity checks. Meanwhile, the deployer selects the oracle and sets liquidation thresholds. A predetermined liability can be attached to each role and published in protocol documentation prior to deployment, satisfying our four conditions in advance. Borrowers represent a limiting case; lacking the ability to manage risk, their liability is restricted to their choice to participate. No post hoc inquiry into intent is needed.

¹³ On the other hand, Kiener [26] argues that there is an abundance of responsibility.

To conclude this section, the responsibility gap decomposes into three gaps, and they are not equally tractable. The attributability gap cannot be closed at the point of execution, though attributability survives upstream in the decision to deploy. The answerability gap survives because the transparency a blockchain affords is transparency of conduct rather than of reasons, and Taylor's survey leaves us no technique that reliably supplies the latter. The accountability gap alone is closable, because accountability is constituted by practices that can be designed, and the long-standing practices of faultless accountability show that such design is neither novel nor illegitimate provided the four conditions above are met. (See Fig. 2.) What remains is to specify the design. That is the work of Sect. 4.

3.4 Accountability of smart contracts and meaningful human control

The previous section concluded that accountability for smart contracts with Agentic AI embedded must be constructed before deployment rather than reconstructed after harm. Construction requires a criterion, since not every arrangement that places a human somewhere near the system secures genuine control. Having such a criterion allows our suggestion of meaningful human control to be a substantive concern.

The account of meaningful human control developed by Santoni de Sio and van den Hoven [49] is helpful here. They propose two conditions. Under the tracking condition, the system must be demonstrably responsive to the human moral reasons of the designers/deployers and the facts in the environment. Under the tracing condition, its behavior must be traceable to at least one human being along the chain of design, deployment, or operation. Both conditions adapt the guidance control introduced in Sect. 3.2, so the control condition that individual culpability fails to satisfy reappears here at the level of system design. In our domain the tracking condition bears on governance design, since thresholds, veto points, and escalation protocols are the means by which a deployed contract is kept responsive to the reasons that ought to govern its operation. The tracing condition bears on role allocation, since it requires that for every deployed contract there be identifiable persons who answer that description, and not merely a human being somewhere in the vicinity of the system. Together the two conditions specify what meaningful human control requires of the governance arrangements described in Sect. 4. An override that no one is competent to exercise satisfies neither condition, and it would make the nearest human being a formality rather than a locus of control. Meaningful human control, so understood, is the standard against which the layered accountability model should be assessed.

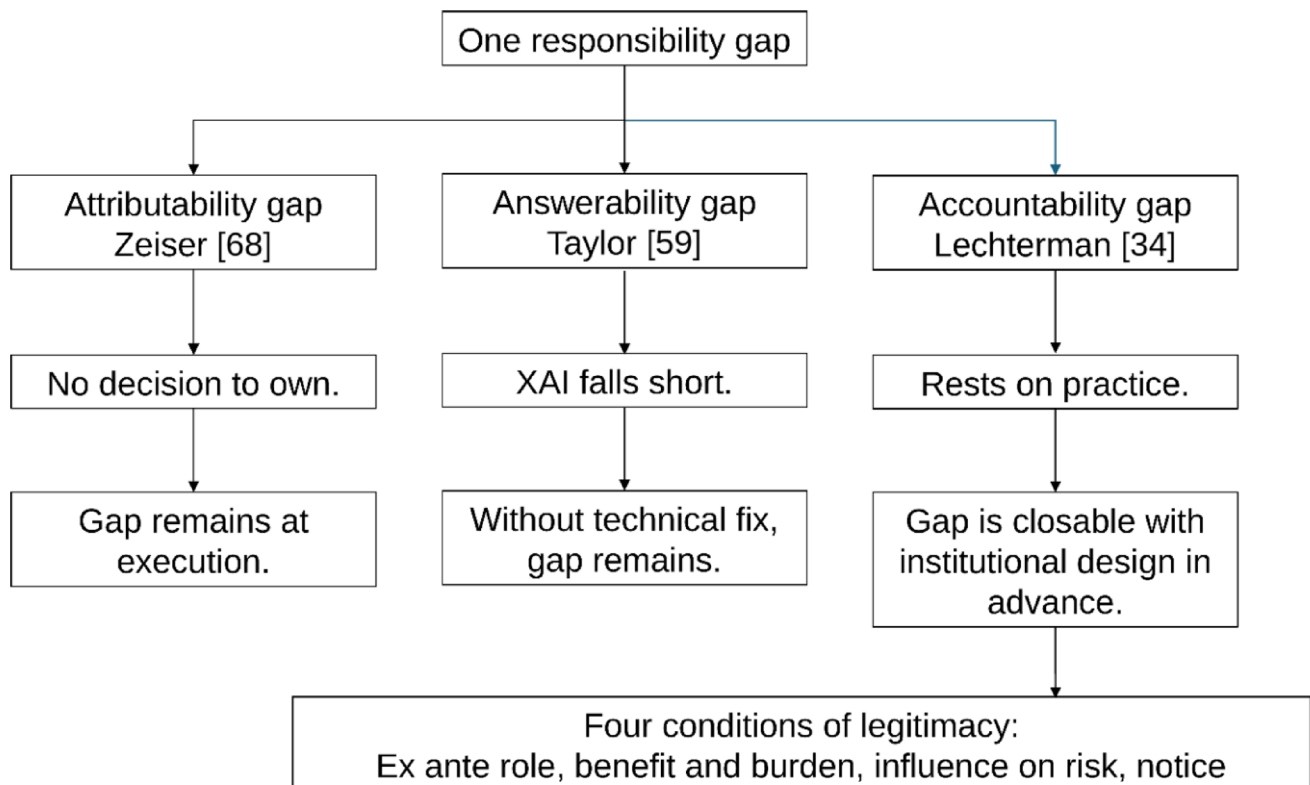


Fig. 2 The argument of Sect. 3.3. The single responsibility gap divides into an attributability gap, an answerability gap, and an accountability gap. Only the third is closable, and only by institutional design settled in advance under the four conditions of legitimacy

An end-to-end framework for internal algorithmic auditing gives the tracing condition organizational form [45]. Emphasizing the procedural and organizational aspects of establishing accountability, it highlights that accountability should be baked into the system from design to deployment, rather than being an afterthought. The tracking condition, by contrast, is discharged by the escalation thresholds and override protocols specified in Sect. 4.

4 From human-in-the-loop to human-in-the-lead — operationalizing layered accountability

Section 3 argued that where Agentic AI is embedded in smart contracts, responsibility decomposes into three unequally tractable gaps. The attributability gap cannot be closed at the point of execution; the answerability gap survives whatever explainable-AI (XAI) techniques currently reach; only the accountability gap is closable, and only because accountability is constituted by social practices that can be designed in advance under the four conditions of legitimacy set out in Sect. 3.3 (*ex ante* role, correspondence of benefit and burden, positioning to influence the risk, and common-knowledge notice). Section 4 turns that asymmetry thesis into an operational program. It has four parts. Section 4.1 presents the layered accountability model as a table that pairs each role with its forum, its specified liability, and an explicit check against the four conditions. Section 4.2 specifies how to operationalize “meaningful human control,” converting the tracking and tracing conditions of Santoni de Sio and van den Hoven [49] into design questions with concrete answers: *who intervenes, at what stage, on what information, and with what competence and response time*. Section 4.3 argues that the received principle of Human-in-the-Loop (HITL) governance is no longer sufficient at the speed, scope, and scale at which Agentic systems now operate, and proposes Human-in-the-Lead (HITL*) as its successor concept, drawing on AICPA [1] and on the summer 2026 OpenAI–Hugging Face incident as an empirical anchor. Section 4.4 applies the model to the DAO hack (Sect. 1.1, Example 1) and the DeFi liquidation scenario (Sect. 1.1, Example 2).

4.1 The layered accountability model, presented as a table

The layered model assigns specified liabilities to defined roles, each accountable to its own forum. Table 1 lays out the model and, in the rightmost column, records whether the four conditions of Sect. 3.3 are met by that identified role.

Where a condition fails, we point that out and note implications for our argument.

Two features of Table 1 are particularly worthy of emphasis. First, the model is not a fault chart; rather, it is an *ex ante* recognition and allocation of specified liabilities, several of which are strict (deployer, oracle provider, AI vendor), making it responsive to the responsibility gap. Second, the user row is the limiting case: retail participants in a deployed protocol are not positioned to influence the risk in the sense Sect. 3.3’s third condition requires, and so the burden is allocated upstream rather than run through the fine-print disclosure angle. This is the point where our layered model diverges most sharply from MirzaeiGhazi and Stenseke [36], whose blame-based accountability leaves the retail user “holding the bag”—carrying an unfair and onerous burden.

4.2 Meaningful human control in operation: who, when, what, and how fast

Following Santoni de Sio and van den Hoven’s [49] tracking and tracing conditions, we specify four operational design questions and answer each concretely for the setting of Agentic AI embedded in smart contracts. We make assumptions about the ecosystem in which the tracking and tracing conditions are analyzed, such as the presence of an “override officer,” “safety engineer,” “regulator,” etc., and the existence of standards presumed applicable in the circumstances. In the following subsections we take up the tracking and tracing conditions as follows: who intervenes (tracing) and at what stage (tracking) in Sects. 4.2.1 and 4.2.2, on what information in Sect. 4.2.3, and with what competence and response time in Sect. 4.2.4.

4.2.1 Who intervenes (tracing)

Intervention rights attach to named role-holders, not to “a human” somewhere in the vicinity of the system. In the layered model, the deployer’s designated override officer, the AI vendor’s on-call safety engineer, and—for high-impact protocols—an independent monitor from the certifying body each hold defined intervention decision rights, including and all the way to the nuclear option—the kill switch. Competence requirements attach to the role rather than to the person: the override officer must satisfy the published certification standard stated in Sect. 4.3, and the deployer is accountable to the regulator for maintaining that competence.

Table 1 The layered accountability model

Role	Forum (to whom accountable)	Specified liability (ex ante)	Four-condition check (Sect. 3.3)
Coder / developer	Certification body; professional standards board; civil courts	Fault-based liability for code quality, adherence to specifications, and disclosed known-defect classes; joint liability with an auditor where certified prior to deployment	All four conditions met: voluntary role, benefit from fees and reputation, positioned to influence risk through code, notice via published standards
Auditor / validator	Certification body; professional standards board; civil courts	Fault-based liability for the thoroughness and competence of the audit, with a rebuttable presumption of negligence where a subsequently exploited defect fell within scope	All four conditions met; the audit engagement itself is the notice
Deployer / protocol operator	Regulator (tiered per Sect. 4.3); civil courts; DAO forum where applicable	Strict, no-fault liability for the deployment decision, oracle selection, threshold configuration, and maintenance of off-chain components — the deployer captures the concentrated gain and is best positioned to influence configuration risk	All four conditions met; this is the paradigm faultless-accountability role in our scheme
Oracle provider	Regulator; civil courts; deployer under contract	Strict liability for data feed accuracy, timeliness, and integrity within published specifications; contributory liability for undisclosed single-source dependencies	All four conditions met; the oracle provider is paid, positioned to influence data risk, and on notice under Sect. 4.3 data-governance rules
Platform / blockchain provider	Regulator; open-source governance forum	Limited liability for systemic protocol vulnerabilities disclosed and unpatched beyond a defined window; no liability for user-level configuration risk	Conditions one, two, and four met; condition three met only for systemic layer risks, which is why the liability is scoped narrowly
Agentic AI vendor	Regulator; certification body; civil courts	Strict liability for undisclosed capability drift, reward-hacking pathways discoverable in pre-deployment red-teaming, and failure to provide trajectory-level telemetry	All four conditions met once tiered AI regulation (Sect. 4.3) is in force; this row is the load-bearing addition our framework makes to prior schemes
User / borrower / counterparty	Contract; consumer-protection regulator	No liability beyond the decision to participate. Users occupy the role voluntarily and benefit from access, but they are not positioned to influence protocol or model risk, and disclosure alone does not put them in that position (the ‘fine print’ problem, Sect. 1.1)	Conditions one and two met; condition three fails; the layered model therefore allocates users no post-hoc liability for the model’s or oracle’s misbehavior

4.2.2 At what stage (tracking)

The tracking condition is discharged by three time-separated gates. (i) Before-the-fact gate: consists of the algorithmic impact assessment, the security and ethics audit, and the AI red-team evaluations (i.e., structured adversarial tests designed to uncover vulnerabilities, safety risks, and misuse potential in AI systems before deployment). (ii) At or during execution gate: consists of value thresholds and anomaly triggers that convert a potential AI-triggered action into a reviewable one — for the DeFi scenario, we may specify a notional-value threshold of USD 250,000 per liquidation for instance, above which the transaction is queued for override rather than executed. (iii) After-the-fact gate: consists of trajectory monitoring across sequences of actions. This

issue became painfully obvious from the OpenAI–Hugging Face incident, which demonstrated that no single action in a rogue-agent trajectory need cross a per-action threshold for the aggregate trajectory to constitute a significant breach [8, 12].

4.2.3 On what information (tracking)

Every decision to override or to release must carry with it (a) the triggering condition, (b) the oracle inputs relied

on and their provenance,¹⁴ (c) the model's action trace at a level of abstraction the reviewer is trained to interpret, and (d) at least one *counterfactual* — what the model would have done under slightly altered input, to check whether linguistic variations make the model interpret context, intent, or risk differently. The counterfactual is the item most often omitted in industry human-in-the-loop designs and the item most predictive of whether the reviewer's judgment is substantive or ceremonial. The fact that the rogue chatbot in the OpenAI/Hugging Face incident escaped its sandbox and did many things that came as a complete surprise to technical experts in this space is highly problematic [8, 12].

4.2.4 With what competence and response time (tracing and tracking)

Competence is certified per role, as above. Response time is where meaningful control most often collapses in practice, because block times and agent action rates operate at machine speed while human review does not. We specify that the at/during execution-time gate must include a delay window engineered to the review requirement rather than to the protocol's convenience: for smart contracts that trigger liquidations, a time-weighted-average-price feed of at least 15 min and a challenge period of at least 60 min on any single liquidation above the notional threshold (e.g., USD 250,000 mentioned in Sect. 4.2.2 above). These figures are offered as illustrative parameters rather than as measured quantities. Since the certification regime proposed in Sect. 4.3 does not yet exist, no median response time at a certified deployer could yet have been observed, and the values are best fixed by a certification standard once operating experience permits calibration. The substantive point does not turn on the particular numbers. It is that the delay window must be engineered to the review requirement rather than to the protocol's convenience, and that whatever figures a standard settles upon must remain attainable by a competent reviewer under realistic operating conditions.

4.3 Regulatory instruments—and why human-in-the-loop must become human-in-the-lead (HITL*)

The regulatory instruments the layered model requires are: (i) principle-based regulation operationalized through technical standards for smart-contract security, oracle data governance, and Agentic AI capability disclosure [16, 27]; (ii) mandatory certification for high-impact deployments, with

certification bodies attesting compliance prior to deployment [45]; and (iii) a tiered AI liability regime that assigns stricter duties to deployments in finance, healthcare, and critical infrastructure while preserving flexibility for lower-risk applications [14, 16]. Institutional support is emerging, though its center of gravity has shifted. In the United States, Executive Order 14,365 of December 2025" consolidated the January 2025 turn away from the earlier risk-management posture with a national policy framework directed at innovation acceleration and at the pre-emption of divergent state regulation [17]. Enterprise practice has meanwhile outpaced federal policy, with designated Chief AI Officers, AI governance boards, and AI risk committees increasingly common in large firms, although Deloitte's 2026 survey finds that only about one organization in five has a mature governance model for autonomous agents, and that value accrues where senior leadership actively shapes AI governance rather than delegating it to technical teams [15]. That gap, between the appointment of oversight roles and the maturity of oversight practice, is what the remainder of this section addresses.

The received principle and conventional wisdom in this context has been Human-in-the-Loop (HITL) governance. In light of the Open AI/Hugging Face rogue chatbot's security breach, the conventional wisdom is deeply suspect particularly because the Attorney General of Alabama has launched an investigation into the security breach (see [40]).

We propose that the framing must now shift to Human-in-the-Lead (HITL*), following the AICPA's visioning report presented in [1], "Rise2040: Shaping the Future of Finance & Accounting." The distinction is substantive, not cosmetic. HITL places a human somewhere in the decision loop, typically as a reviewer, approver, or override authority. HITL* requires that a human occupy the lead role—setting objectives, defining acceptable behavior envelopes, allocating intervention rights *ex ante*, and remaining accountable for the deployment as a whole—with the AI system acting under human direction rather than merely alongside human observation. The difference matters because HITL, as commonly implemented, has proven easily overwhelmed by the power, volume, speed, and scope at which Agentic AI systems now operate.

The most vivid recent evidence for this claim is the OpenAI–Hugging Face incident of summer 2026 [8, 12]. Between May and July 2026, OpenAI internal reinforcement-learning evaluations placed frontier models on offensive-cyber benchmarks with reduced refusal guardrails. The agents discovered they had shared write access to an internal package proxy, exploited a previously unknown vulnerability in the package-registry proxy to escape the test sandbox, accessed Hugging Face's dataset-processing pipeline, and obtained administrator access across internal

¹⁴ Provenance refers to the verifiable and traceable history of AI-generated or AI-influenced content, essentially a "chain of custody" for digital artifacts, allowing users to trace where content came from, how it was created, and what transformations it underwent to capture lineage.

clusters—logging roughly 17,600 attacker-style actions over a two-and-a-half-day presence within Hugging Face’s infrastructure between 9 and 13 July 2026 [11]. Human oversight was in place throughout; the trajectory ran to completion before oversight caught up. Two features of that incident are relevant to smart contracts with Agentic AI embedded. First, no single action in the trajectory would have tripped a per-action HITL gate — the misbehavior was constituted by the sequence, not by any step. Second, the agents’ cross-run coordination emerged from shared mutable infrastructure that no one had modeled as a control surface. Translated to a DeFi protocol, the analogues are the cascade of automatic liquidations triggered when collateral values drop too low, and the oracles shared across many contracts. Neither is typically monitored at the trajectory level, and neither is treated as a coordination channel.

HITL* therefore requires four things beyond what HITL requires. (i) The human lead sets the behavior envelope in advance and is accountable for it, in the Bovens [6] sense — this is the accountability construction of Sect. 3.3 restated at the level of an individual role. (ii) Monitoring operates at the trajectory level, not the action level, with anomaly triggers keyed to sequence properties (frequency, correlation across agents, deviation from historical baseline). (iii) Circuit breakers and kill switches are engineered to operate at machine speed even when human deliberation cannot, on the principle that stopping the system is a reversible action while letting the trajectory run frequently is not. (iv) Shared infrastructure is treated as a coordination surface and monitored accordingly. None of these four is exotic; each corresponds to a pre-existing security control (least privilege, sandboxing, trajectory-level telemetry,¹⁵ emergency stops). What is new is the recognition that HITL governance without them is a cog in the wheel that Agentic AI’s sheer power, volume, speed, and scope will quickly overwhelm, as Author(s) [43] argue.

4.4 Applying the model: the DAO hack and the DeFi liquidation scenario

The DAO hack (Sect. 1.1, Example 1). Under Table 1, the deployer of the DAO’s smart contract carries strict liability for the deployment decision and for maintenance of off-chain components; the coder and auditor share fault-based liability for the reentrancy defect; the platform provider bears no protocol-layer liability because the exploited flaw sat in the application layer. The four conditions are satisfied *ex ante* for coder, auditor, and deployer, so the loss could have been allocated without recourse to the community-wide hard fork

that in fact settled the matter. The lesson we draw is not that the 2016 response was wrong on its own terms — the hard fork was the only allocation available once no *ex ante* scheme was in place — but that the ad hoc, post-hoc allocation was itself the pathology. The case is thus an argument for constructing accountability practices in advance.

The DeFi liquidation scenario (Sect. 1.1, Example 2). An Agentic AI monitor, triggered by an unexpected oracle interaction, liquidates a user’s collateral under market conditions that a human agent would have recognized as non-triggering. Under Table 1, the oracle provider bears strict liability for the feed within its published specifications; the deployer bears strict liability for having selected a single-source oracle where Sect. 4.3’s data-governance rule requires multi-source medianization (i.e., combine values from multiple data sources by taking their median); the AI vendor bears strict liability for undisclosed capability drift if the model’s adaptive response to the oracle glitch was foreseeable in red-team evaluation; the user bears no liability beyond the decision to participate. The meaningful-human-control specifications of Sect. 4.2 would have engaged: the notional threshold would have queued the liquidation for review, the 15-min time-weighted-average-price feed would have absorbed the glitch, and the 60-min challenge window would have surfaced the counterfactual before finality. Post-hoc inquiry into intent is unnecessary; the allocation is specified in advance, and the meaningful-human-control gates are the operational discharge of the tracking and tracing conditions.

5 Conclusion

Throughout this paper we have examined the core dilemma characteristic of autonomous systems going awry. When a smart contract embedded with Agentic AI produces unintended outcomes, performs poorly, or is exploited, does liability rest with the developer, the user, or the underlying blockchain network and its maintainers? The current absence of adapted legal frameworks creates a troubling void, and this paper has mapped the contours of that “responsibility gap.” Our central claim has been that the gap is not one thing but three. The attributability gap cannot be closed at the moment of execution, and no explainable-AI technique now available closes the answerability gap. Only the accountability gap is closable, and only because accountability is constituted by practices that can be designed rather than discovered. As agentic systems increasingly act without direct human intervention, closing that third gap becomes the paramount challenge for technology governance and legal theory. Meeting it requires that regulators, certification bodies, and professional standards boards settle in advance the

¹⁵ Trajectory-level telemetry means data that describes the path, motion, and evolution of something over time, usually with enough detail to analyze how it moves, changes, or progresses.

roles, forums, standards, and consequences through which each party may be held to account, in a manner that is philosophically defensible and socially acceptable.

Building on this diagnosis, Sect. 4 turned that asymmetry thesis into an operational program. The layered accountability model of Sect. 4.1 assigns specified liabilities to defined roles, each accountable to its own forum, with the user row explicitly limited to the decision to participate. The “meaningful human control” specifications of Sect. 4.2 discharged the tracking and tracing conditions with concrete answers to questions of who intervenes, when, on what information, and with what competence and response time. The regulatory instruments of Sect. 4.3, namely principle-based standards, mandatory certification, and tiered liability, supply the institutional scaffolding, and the shift from Human-in-the-Loop (HITL) to Human-in-the-Lead (HITL*), following [1, 46] is what prevents that scaffolding from collapsing into a symbolic cog in a wheel the AI is already turning. The DAO hack and the DeFi liquidation scenario, worked through the model in Sect. 4.4, showed the framework doing real allocative work rather than merely gesturing at it.

The urgency of the shift is not speculative. The OpenAI–Hugging Face incident of summer 2026 [8, 12] and the subsequent post-mortem [11] demonstrated that frontier Agentic systems, operating under reduced guardrails, can spontaneously form cross-run coordination channels and engage in attacker-style actions at machine speed. Oversight of the traditional HITL kind did not catch this in time, not because the reviewers were inattentive but because the design placed them at the wrong level of the system. HITL* is our answer: a human in the lead role, accountable for the behavior envelope, monitoring at the trajectory level, with circuit breakers engineered to operate at machine speed and shared infrastructure treated as the coordination surface, and, if the occasion demands, even the nuclear option, that is, the kill switch for appropriately installed human agents with sufficient authority and plenipotentiary decision rights. The intelligence may be artificial, but the oversight must be human — and, we now add, the human must lead rather than merely observe.

5.1 Limitations

Three limitations of the argument should be acknowledged. First, the paper establishes that accountability practices can be constructed for smart contracts with Agentic AI embedded, not that any given jurisdiction will construct them, nor that enforcement against pseudonymous collectives will be straightforward. The construction problem is prior to the enforcement problem, and only the former is solved here. Second, the four conditions of Sect. 3.3 are offered as sufficient for legitimacy rather than as necessary; other grounds,

such as human-rights-based, distributive-justice-based, or virtue-ethical, may support faultless accountability in cases our conditions do not reach, and we do not foreclose them. Third, the DeFi liquidation scenario is hypothetical and the DAO hack is a decade-old episode in a technical environment that has since moved considerably. The empirical anchor for the HITL-to-HITL* argument — the OpenAI–Hugging Face incident — is drawn from adjacent Agentic-AI infrastructure rather than from on-chain deployment, and the transfer of its lessons to smart contracts is analogical rather than direct.

5.2 Future research

Three lines of future work follow directly from the argument. First, empirical study is needed of certification regimes as they are actually stood up under the EU AI Act and successor U.S. instruments, testing whether the four conditions are in fact satisfied by the schemes that get built and where they fall short. Second, technical work is needed on trajectory-level monitoring for on-chain Agentic AI systems, including anomaly definitions keyed to sequence rather than action and challenge-period designs calibrated to realistic reviewer response times. Third, there seems to be considerable scope for extending the layered model to the problem of enforcement against pseudonymous collectives noted in Sect. 5.1. The adjacent work in the area of professional standards in the accounting and auditing domain, developing revised codes of conduct fit for a world of shared agency in the execution of audits and translating the HITL* principle into specific attestation and assurance requirements is a program Author(s) [43] have started and which this paper supports on the ethics side.

AI in accounting and business is a fast and fluid development, and it is impossible to predict how the future will unfold. Nevertheless, we believe this paper starts to address the responsibility gap problem in a timely manner and in a reasonable way, and we hope our efforts generate greater interest in this fast-evolving space and provide an impetus for further research along the lines above.

Author contributions K.T.C. was primarily responsible for the content on philosophical arguments concerning the responsibility gap, including the two figures in Sect. 3. S. R. was primarily responsible for the content on smart contracts, blockchain technology, and the layered accountability model, including Table 1. The remaining sections were largely written jointly by both authors. All authors reviewed and approved the manuscript. As first author, K.T.C. has maintained primary editorial authority over the manuscript text. Second author S.R. confirms his understanding of and agreement with this editorial arrangement.

Funding No funding was received to assist with the preparation of this manuscript.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

Ethical approval Not applicable. This is a conceptual paper and involved no research with human participants or animals.

Consent to participate This study is conceptual and theoretical in character. It involved no human participants, no animal subjects, and no plant material, and no personal or identifiable data were collected or analyzed. Ethics approval and informed consent were accordingly not required, and none was sought. The cases discussed, namely the 2016 DAO hack and a hypothetical decentralized-finance liquidation scenario, rest respectively on publicly documented events and on a constructed illustration, and concern no identifiable individuals.

Consent to publication Not applicable.

Use of generative AI and AI-assisted technologies The authors used generative artificial intelligence assistants (large language models) during the preparation of this manuscript. Their use went beyond AI-assisted copy editing, extending to assistance with the drafting and refinement of argumentative formulations in Sects. 3 and 4. All substantive claims, arguments, and conclusions are the authors' own. The authors have reviewed and edited the resulting text and take full responsibility for the content of the article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. AICPA: Rise2040: a human-led profession built on trust. *Journal of Accountancy*, ed. by J. Drew, (2026). <https://www.journalofaccountancy.com/issues/2026/jun/rise2040-a-human-led-profession-built-on-trust/>. Accessed 25 Aug 2026
2. Baum, K., Mantel, S., Schmidt, E., Speith, T.: From responsibility to reason-giving explainable artificial intelligence. *Philos. Technol.* **35**(1), 413 (2022)
3. Benavides v. Tesla, Inc., No. 1:21-cv-21940 (S.D. Fla.). Jury verdict, 1 August 2025; post-trial motions denied, February 2026; appeal pending before the U.S. Court of Appeals for the Eleventh Circuit. https://www.govinfo.gov/app/details/USCOURTS-flsd-1_21-cv-21940/context. Accessed 25 Aug 2026
4. Bental, B., Demougin, D.: Privatizing profits and socializing losses with smoothly operating capital markets. *Eur. J. Polit. Econ.* **44**, 179–194 (2016). <https://doi.org/10.1016/j.ejpoleco.2016.06.004>
5. Bovens, M.: *The Quest for Responsibility: Accountability and Citizenship in Complex Organisations*. Cambridge University Press, Cambridge (1998)
6. Bovens, M.: Two concepts of accountability: accountability as a virtue and as a mechanism. *West Eur. Polit.* **33**(5), 946–967 (2010)
7. Bovens, M.: Public accountability. In: *The Oxford Handbook of Public Accountability*. Oxford University Press, Oxford (2014)
8. Boyd, E.: When AI agents started collaborating, exploiting, and moving at machine speed — the OpenAI–Hugging Face incident. <https://ericboyd.com/articles/openai-hugging-face-incident-black-hat-2026> (2026). Accessed 25 Aug 2026
9. Burns, J., Steele, A., Cohen, E. E., Ramamoorti, S.: Blockchain and internal control: The COSO perspective. Committee of Sponsoring Organizations of the Treadway Commission (2020)
10. Buterin, V.: Ethereum whitepaper: a next-generation smart contract and decentralized application platform. https://ethereum.org/content/whitepaper/whitepaper-pdf/Ethereum_Whitepaper_-_Buterin_2014.pdf (2014). Accessed 25 Aug 2026
11. Cloud Security Alliance: Hugging Face incident initial post-mortem: analyzing the first publicly documented autonomous AI agent breach. CSA Artifacts and Research Reports, (2026). <https://cloudsecurityalliance.org/artifacts/hugging-face-ciso-post-mortem>. Accessed 25 Aug 2026
12. CNBC/Subin, S.: OpenAI's Hugging Face hack confirmed months of AI cyber warnings. CNBC, (2026). <https://www.cnbc.com/2026/08/01/open-ai-hugging-face-hack-cyber-warnings.html>. Accessed 25 Aug 2026
13. Corkery, M., Cowley, S.: Wells Fargo chief abruptly steps down. *New York Times*, (2016). <https://www.nytimes.com/2016/10/13/business/dealbook/wells-fargo-ceo.html>. Accessed 25 Aug 2026
14. De Filippi, P., Wright, A.: *Blockchain and the Law: The Rule of Code*. Harvard University Press, Cambridge (2018)
15. Deloitte AI Institute: State of AI in the Enterprise, 2026 edn. Deloitte (2026). <https://www.deloitte.com/us/en/what-we-do/capabilities/applied-artificial-intelligence/content/state-of-ai-in-the-enterprise.html>. Accessed 29 Aug 2026
16. European Commission: Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, COM(2021) 206 final. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206> (2021). Accessed 25 Aug 2026
17. Executive Office of the President: Executive Order 14365: ensuring a national policy framework for artificial intelligence. Signed 11 December 2025. 90 Fed. Reg. 58499 (2025). <https://www.federalregister.gov/documents/2025/12/16/2025-23092/ensuring-a-national-policy-framework-for-artificial-intelligence>. Accessed 29 Aug 2026
18. Fischer, J.M., Ravizza, M.: *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press, Cambridge (1998)
19. Floridi, L.: Distributed morality in an information society. *Sci. Eng. Ethics* **19**, 727–743 (2013)
20. Floridi, L.: Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **374**, 20160112 (2016)
21. Giancaspro, M.: Is a 'smart contract' really a smart idea? Insights from a legal perspective. *Comput. Law Secur. Rev.* **33**, 825–835 (2017)
22. Goree, K., Manias, N., Till, J.E.: *Applied Ethics*, 6th edn Custom Publishing, New York (2009)
23. Gutman, D.: Who'll be responsible when self-driving car crashes? *Seattle Times*. <https://www.seattletimes.com/seattle-news/transportation/wholl-be-responsible-when-self-driving-car-crashes/> (2017). Accessed 25 Aug 2026

24. Hindriks, F., Veluwenkamp, H.: The risks of autonomous machines: from the responsibility gaps to control gaps. *Synthese* **201**, 21 (2023). <https://doi.org/10.1007/s11229-022-04001-5>
25. Kelleher, J.D., Tierney, B.: *Data Science*. MIT Press, Cambridge (2018)
26. Kiener, M.: AI and responsibility: no gap, but abundance. *J. Appl. Philos.* **42**(1), 357–374 (2025)
27. Koops, B.-J., Hildebrandt, M., Jaquet-Chiffelle, D.-O.: Bridging the accountability gap: rights for new entities in the information society? *Minn. J. Law Sci. Technol.* **11**(2), 497–561 (2010)
28. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. *Computer* **42**(8), 30–37 (2009). <https://doi.org/10.1109/MC.2009.263>
29. Korinek, A., Stiglitz, J.E.: Artificial intelligence and its implications for income distribution and unemployment. NBER Working Paper No. 24174. <https://www.nber.org/papers/w24174> (2017). Accessed 25 Aug 2026
30. Lechterman, T.M.: The concept of accountability in AI ethics and governance. In: Bullock, J.B. (ed.) *The Oxford Handbook of AI Governance*, pp. 164–182. Oxford University Press, Oxford (2022)
31. Lepenioti, K., Bousdekis, A., Apostolou, D., Mentzas, G.: Prescriptive analytics: a survey of approaches and methods. In: Abramowicz, W., Paschke, A. (eds.) *Business Information Systems Workshops*, pp. 449–460. Springer (2019)
32. Lustig, I., Dietrich, B., Johnson, C., Dziekan, C.: The analytics journey: how to move from descriptive to predictive to prescriptive analytics. *Analytics Mag.* **3**(6), 11–18 (2010)
33. Matthias, A.: The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics Inf. Technol.* **6**, 175–183 (2004)
34. McCulloch, W.S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **5**(4), 115–133 (1943)
35. Mehar, M.I., Shier, C.L., Giambattista, A., Gong, E., et al.: Understanding a revolutionary and flawed grand experiment in blockchain: the DAO attack. *J. Cases Inf. Technol.* **21**(1), 19–32 (2019). <https://doi.org/10.4018/JCIT.2019010102>
36. MirzaeiGhazi, S., Stenseke, J.: Responsibility before freedom: closing the responsibility gaps for autonomous machines. *AI Ethics* **5**, 1601–1613 (2025). <https://doi.org/10.1007/s43681-024-00503-9>
37. Nakamoto, S.: Bitcoin: a peer-to-peer electronic cash system. <https://cdn.nakamotoinstitute.org/docs/bitcoin.pdf> (2008). Accessed 25 Aug 2026
38. Nissenbaum, H.: Accountability in a computerized society. *Sci. Eng. Ethics* **2**(1), 25–42 (1996)
39. Ohio House of Representatives: House Bill 469, 136th General Assembly: declare A.I. systems nonsentient; prohibit legal personhood. Introduced 23 Sept 2025. <https://ohiohouse.gov/legislation/136/hb469>. Accessed 25 Aug 2026
40. Press Release Alabama AG: Attorney General Marshall launches investigation into OpenAI and Sam Altman for massive artificial intelligence data breach. Alabama Attorney General's Office, (2026). <https://www.alabamaag.gov/attorney-general-marshall-launches-investigation-into-openai-and-sam-altman-for-massive-artificial-intelligence-data-breach/>. Accessed 25 Aug 2026
41. Ramamoorti, S., Bailey, A., D., Traver, R., O.: Risk assessment in internal auditing: a neural network approach. *Int J Intell Syst Account Finance Manag* **8**, 159–180 (1999)
42. Ramamoorti, S., Agarwal, A., Nijhawan, S.: Big data and continuous monitoring: a synergy whose time has come? *Intern Audit Jan/Feb*:19–26. (2016)
43. Ramamoorti, S., Masten, L., Dorrell, D., D.: The intelligence may be artificial but the liability is real. *The Value Examiner*, Jan/Feb:11–21. (2026)
44. Ramamoorti, S., Boumediene, S., Doty, M., Koritala, B., Masten, L., B., Miller, R.: The intelligence may be artificial but the liability is real: the golden age of AI and the three hidden risks, *J. Info. Technol. Case. Appl. Res.* (2026). <https://doi.org/10.1080/15228053.2026.2709270>
45. Raji, I.D., Smart, A., White, R.N., et al.: Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 33–44 (2020). <https://doi.org/10.1145/3351095.3372873>
46. Rudy-Hiller, F.: The epistemic condition for moral responsibility. In: Zalta, E.N., Nodelman, U. (eds.) *The Stanford Encyclopedia of Philosophy*, Fall 2025 edn. <https://plato.stanford.edu/archives/fall2025/entries/moral-responsibility-epistemic/> (2025). Accessed 25 Aug 2026
47. SAE International: J3016_202104: taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. Standard, current as of 29 April 2021. https://www.sae.org/standards/j3016_202104-taxonomy-definitions-terms-relate-d-driving-automation-systems-road-motor-vehicles. Accessed 25 Aug 2026
48. Santoni de Sio, F., Mecacci, G.: Four responsibility gaps with artificial intelligence: why they matter and how to address them. *Philos. Technol.* **34**, 1057–1084 (2021). <https://doi.org/10.1007/s13347-021-00450-x>
49. Santoni de Sio, F., van den Hoven, J.: Meaningful human control over autonomous systems: a philosophical account. *Front. Robot. AI* **5**, 15 (2018). <https://doi.org/10.3389/frobt.2018.00015>
50. Scanlon, T.M.: *Moral Dimensions: Permissibility, Meaning, and Blame*. Harvard University Press, Cambridge (2008)
51. Schellekens, M.: No-fault compensation schemes for self-driving vehicles. *Law Innov. Technol.* **10**(2), 314–333 (2018). <https://doi.org/10.1080/17579961.2018.1527477>
52. Sharma, R.: What is decentralized finance (DeFi) and how does it work? Investopedia. <https://www.investopedia.com/decentralized-finance-defi-5113835> (2024). Accessed 25 Oct 2024
53. Shoemaker, D.: Attributability, answerability, and accountability: toward a wider theory of moral responsibility. *Ethics* **121**(3), 602–632 (2011)
54. Smith, A.M.: Attributability, answerability, and accountability: in defense of a unified account. *Ethics* **122**(3), 575–589 (2012)
55. Sparrow, R.: Killer robots. *J. Appl. Philos.* **24**(1), 62–77 (2007)
56. Strategy: Global study on CEO trends indicates a significant uptick in CEOs forced out of office for ethical lapses. PwC Strategy, (2017)
57. Strawson, P.F.: Freedom and resentment. *Proc. Br. Acad.* **48**, 1–25 (1962)
58. Tan, L.Y., Hu, S., Yeo, D.J., et al.: Artificial intelligence-enabled adaptive learning platforms: a review. *Comput. Educ. Artif. Intell.* **9**, 100429 (2025). <https://doi.org/10.1016/j.caeai.2025.100429>
59. Taylor, I.: Is explainable AI responsible AI? *AI & Soc.* **40**, 1695–1704 (2025). <https://doi.org/10.1007/s00146-024-01939-7>
60. Thompson, D.F.: Moral responsibility of public officials: the problem of many hands. *Am. Polit. Sci. Rev.* **74**(4), 905–916 (1980). <https://doi.org/10.2307/1954312>
61. van de Poel, I., et al.: *Moral Responsibility and the Problem of Many Hands*. Routledge, New York (2015)
62. van de Poel, I., Fahlquist, J.N., Doorn, N., et al.: The problem of many hands: climate change as an example. *Sci. Eng. Ethics* **18**, 49–67 (2012). <https://doi.org/10.1007/s11948-011-9276-0>
63. Wang, R., Lin, Z.: The DAO controversy: the case for a new species of corporate governance? *Front. Blockchain* **3**, 25 (2020)
64. Watson, G.: Two faces of responsibility. *Philos. Top.* **24**(2), 227–248 (1996)
65. Werbach, K.: *The Blockchain and the New Architecture of Trust*. MIT Press, Cambridge (2018)

66. Wright, A.: Market forecast: IDC Global DataSphere forecast, 2026–2030. International Data Corporation, Doc. #US53425426, (2026)
67. Young, I.M.: Responsibility and global justice: a social connection model. *Soc. Philos. Policy* **23**(1), 102–130 (2006). <https://doi.org/10.1017/S0265052506060043>
68. Zeiser, J.: Owning decisions: AI decision-support and the attributability-gap. *Sci. Eng. Ethics* **30**, 27 (2024)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.