

THE RESPONSIBILITY GAP IN EXPLAINABLE AI

Filippos Stamatiou  filippostam@gmail.com

Centre for the Philosophy of Artificial Intelligence, University of Copenhagen, Denmark

Unit for the Ethics of Technology, Stellenbosch University, South Africa

Article type: Research article

Review process: Double-blind peer review

Topical collection: Ethics and Normativity of Explainable AI: Explainability as a Social Practice, guest editors: Suzana Alpsancar, Tobias Matzner, Wessel Reijers

This open-access article is published with a Creative Commons CC-BY 4.0



license

<https://creativecommons.org/licenses/by/4.0/>

DOI: <https://doi.org/10.59490/jhtr.2026.4.8241>

ISSN: 2773-2266

Submitted: 30 June 2025 **Revised:** 09 March 2026, 26 May 2026 **Accepted:** 16 June 2026 **Published:** 13 October 2026

How to cite (APA): Stamatiou, F. (2026). The responsibility gap in explainable AI. *Journal of Human-Technology Relations*, 4, 1-26. <https://doi.org/10.59490/jhtr.2026.4.8241>

Corresponding author: Filippos Stamatiou

©2026 The Author, published by TU Delft OPEN Publishing on behalf of the authors.

Keywords

Explainable AI; Responsibility Gap; Moral Responsibility; XAI; Mechanistic Interpretability; Accountability

Abstract

This paper critically examines the persistent responsibility gap in explainable artificial intelligence (XAI), arguing that current approaches fail to adequately bridge it. While XAI is often promoted as a remedy for the opacity of AI systems—the so-called “black box problem”—and as a way to enable human accountability in both high- and low-stakes AI applications, I contend that these hopes are premature. The paper distinguishes between legal and moral responsibility, arguing that regulation alone cannot address the full spectrum of responsibility gaps, especially in everyday, low-stakes scenarios. Then, I argue that XAI needs to tackle two separate but interconnected problems in its attempt to bridge the responsibility gap: an epistemological problem concerning the nature of XAI explanations, and a moral attribution problem regarding the distribution of blame and praise in situations where AI is involved in decision-making. Through detailed discussion of XAI methods and their limitations, I demonstrate that existing techniques often provide only post-hoc, technically focused explanations, which are insufficient for grounding moral responsibility. The paper concludes that without fundamental advances in both the design of explanations and our practices of responsibility attribution, the responsibility gap in AI will remain an unresolved and pressing issue for the field.

Plain Language Summary¹

- This paper is relevant to anyone who uses artificial intelligence (AI), as it asks who should take responsibility when an AI system causes harm. Problems can arise even in everyday situations, such as an email filter hiding a booking confirmation and prompting someone to pay twice.
- The author argues that explaining an AI decision does not automatically make anyone accountable for it. Two challenges must be addressed together: helping people understand how the system works and deciding how responsibility should be shared among those who develop, provide, and use it.
- Many current explanation tools produce simplified accounts of AI decisions after the fact. These can create a false sense of understanding. For instance, a radiologist may fully grasp why an AI flagged a scan as abnormal yet still feel unable to override it in practice due to institutional pressure or time constraints. So, understanding what a system did and having real power to act differently are two distinct things, and existing tools rarely close that gap.
- The article proposes two kinds of reform. The first is technical: explanations should be redesigned to flag ethical concerns directly, not just which features drove a decision, but whether those features reflect fairness or bias. The second draws on Aristotle's concept of practical wisdom (phronesis): responsible AI use is a skill that must be cultivated through practice, not just a matter of having the right information available.
- A distinctive contribution is the connection between better AI explanations and a broader rethink of responsibility. Drawing on Ubuntu, an African philosophical tradition that emphasises human interconnectedness, the author proposes shared efforts to prevent harm rather than focusing solely on individual blame after something goes wrong.

¹ AI-generated; author checked and approved.

- These arguments matter beyond academic philosophy. Laws and regulations can only reach so far — everyday AI interactions, from spam filters to navigation apps, fall outside most legal frameworks but still produce real harms with no clear answer as to who should bear responsibility. The paper argues that closing this gap requires both better-designed explanations and a genuine rethink of how responsibility is shared as a precondition for any regulation to be meaningful.

1 INTRODUCTION

Artificial intelligence is increasingly present in our lives, ranging from high-stakes applications like healthcare and criminal justice to commonplace applications such as email spam filtering. The growing complexity of machine learning systems creates opacity that even their designers cannot always penetrate, known as the black box problem (Pasquale, 2015; Rudin & Radin, 2019). This epistemic barrier impedes not only technical understanding but also accountability, fueling the emergence of responsibility gaps—situations where it is unclear or potentially impossible to determine who should be held accountable for AI-related harms (Matthias, 2004; Nyholm, 2023).

The distinction between legal and moral responsibility is critical: legal frameworks strive to assign formal blame according to established principles but may fall short in addressing the nuanced, informal harms AI systems can produce in daily human interactions (Bathae, 2018). While recent regulatory efforts, such as the EU AI Act (Regulation (EU) 2024/1689), attempt to bridge these gaps in high-stakes domains, low-stakes uses, like the spam filter case, often remain impervious to legal remedies, highlighting the persistent relevance of moral responsibility gaps. This paper argues that explainable artificial intelligence (XAI) has been proposed as a remedy for the opacity of AI systems and as a means to provide human accountability. Yet, these hopes are premature. The responsibility gap remains, requiring advances not only in explanations and interpretability, but also in our practices for distributing and attributing both legal and moral responsibility, as recent literature has shown (Hindriks & Veluwenkamp, 2023; Conradie, 2023; Nyholm, 2018).

In the following sections, the conceptual background will be outlined, including new regulatory developments, recent advances in responsibility gap scholarship, and a clarified distinction between epistemological and moral problems. The overall aim is to provide an integrated account of why the responsibility gap persists and what changes are needed so XAI can truly help bridge it.

2 RESPONSIBILITY GAPS

2.1 THE BLACK BOX PROBLEM

The “black box problem” in artificial intelligence refers to the opacity of machine-learning systems, particularly those employing deep neural networks and other complex algorithms, which resist human comprehension despite their decision-making and predictive efficacy (Pasquale, 2015; Rudin & Radin, 2019). This problem is essentially an epistemic barrier to understanding AI, resulting from two interrelated technical factors: the distributed complexity of deep learning architectures and the high-dimensional geometric transformations in models like support vector machines. These systems process inputs through layers of non-linear computations or map data into abstract mathematical spaces, rendering their internal logic largely inaccessible to intuitive or analytical scrutiny (Burrell, 2016; Zednik, 2021).

The opacity of AI systems exposes long-standing tensions between predictive accuracy and explanatory depth. Why do we require AI systems that are transparent if they already provide superior accuracy? This debate evokes essential considerations about informed consent and due process in AI applications. Furthermore, the black box problem disrupts two distinct foundational elements of human-AI interaction: our conceptual reliance on intentionality (understood as purposeful mental states) and our practical need for causal explanation (understood as traceable chains of events leading to outcomes).

Existing legal frameworks become inadequate for AI systems in two corresponding ways. First, they depend on discerning intent (typically human mental states like purpose or foresight), which AI lacks in any meaningful sense (Bathae, 2018). Second, they require establishing proximate causation through reconstructible decision pathways, which becomes impossible when AI internals defy human tracing due to their complexity (Bathae, 2018; Pasquale, 2015). This challenge is especially pronounced in high-stakes domains such as healthcare and criminal justice, where opaque models have come under pressure for undermining human control and accountability (Burrell, 2016; Zednik, 2021; Hindriks & Veluwenkamp, 2023).

A few clarifications regarding the black box problem are in order. First, while it is ultimately the outputs of AI models that matter to us, the challenge arises precisely because the internal processes generating those outputs are too complex for human comprehension, even for their designers (Burrell, 2016; Rudin & Radin, 2019). As these outputs become increasingly consequential in both high- and low-stakes applications, it is imperative to have reliable explanations of the decision processes that produced them. Second, the problem is not merely that users lack understanding of certain outputs, though this is significant. More fundamentally, even the designers and engineers who create these models often cannot fully comprehend their internal workings due to their distributed complexity and high-dimensional transformations. Using technology without understanding it fully is both common and generally unproblematic; consider smartphones or modern cars. But using technology that no one, not even its creators, fully understands represents a distinct epistemic challenge with serious accountability implications.

2.2 RESPONSIBILITY GAPS

Responsibility gaps are situations where it is unclear or, according to some accounts, in principle impossible to identify morally appropriate parties to hold accountable for harms caused by AI systems, particularly when those systems are opaque or poorly understood (Matthias, 2004;

Nyholm, 2023). While some theorists emphasize genuine accountability gaps where no suitable responsible agent exists due to the inherent features of automation (Matthias, 2004), others argue such gaps may be resolvable through institutional or conceptual innovation (Tigard, 2021). Responsibility gaps have been linked to diverse AI applications, including military technologies (Sparrow, 2007), healthcare (Lang et al., 2023), finance (Buckley et al., 2021; Zetzsche et al., 2020; Ridzuan et al., 2024), and legal systems (Danaher, 2016). Some scholars warn that responsibility gaps threaten the foundations of societal moral frameworks and legal liability (Matthias, 2004), while others are sceptical about their universal significance (Tigard, 2021).

The emergence of responsibility gaps in AI has sparked a range of responses. Some advocate regulatory actions, such as banning high-risk applications like lethal autonomous weapons (Sparrow, 2007). Others engage in theoretical and practical efforts to accept (Coeckelbergh, 2020; Oimann, 2023), overcome (Tigard, 2021; Vallor & Vierkant, 2024), deny (MirzaeiGhazi & Stenseke, 2025; Königs, 2022; Hindriks & Veluwenkamp, 2023), or bridge these gaps by reimagining moral and legal responsibility frameworks and governance mechanisms (Danaher, 2022; Munch et al., 2023; Königs, 2022; Nyholm, 2018; Kiener, 2022).

At the heart of responsibility gaps lie the opacity and complexity of AI systems, which obscure causal pathways and reduce stakeholders' capacity for control. This results in moral and legal ambiguity, complicating the attribution of accountability. Responsibility gaps call for a rethinking of traditional accountability models in light of growing autonomous and opaque decision-making systems.

2.3 BRIDGING RESPONSIBILITY GAPS

A promising way to respond to responsibility gaps is by regulating the design, deployment, and use of certain technologies. When a new technology exhibits potential for unforeseeable harmful consequences, restricting its applications seems the safest way forward. Nuclear energy is a good historical example. While the benefits of nuclear reactors to produce energy through atomic fission are significant, only a handful of countries have operating nuclear power stations, and these are subject to strict monitoring and regulation. The dangers of unrestricted nuclear energy outweigh the benefits, necessitating regulation to reduce the likelihood of accidents.

Perhaps regulation is the appropriate response to AI-related responsibility gaps. This response functions on two levels: one involving developers and the tech sector, and the other involving the public and regulatory institutions. First, AI encompasses a broad and dynamic range of technologies that are gradually better understood. As we improve training and understanding of AI models, their potential implications become more foreseeable. A more mature tech sector can predict what a model may do and thereby accept responsibility for some outcomes while not for others, minimizing undesirable consequences. Second, as AI applications become increasingly ubiquitous, the public grows more familiar with their uses and potential misuse. A more mature public can more effectively regulate AI's harmful applications. Legislation governing the development and deployment of potentially harmful AI systems is inevitable (Oimann, 2023; Regulation (EU) 2024/1689).

Between a private tech sector capable of foreseeing most responsibility gaps before they occur and a public sector empowered to enact regulation for the remainder, there exists an optimistic scenario where responsibility gaps gradually decrease. Consider the paradigmatic case of an autonomous vehicle injuring a pedestrian. Is the car manufacturer to blame? Or the software developer? Or does the passenger implicitly accept moral implications by choosing to ride in the autonomous vehicle, notwithstanding risks beyond their control? Similar scenarios unfolding in fields such as autonomous weapons (Oimann, 2023), healthcare, finance, or legal applications

(Buckley et al., 2021) may be addressed through regulation, such as the EU Artificial Intelligence Act designed to manage high-stakes AI risks.

However, regardless of this optimistic projection, regulation alone cannot close all responsibility gaps. If regulation were sufficient to bridge responsibility gaps, all such gaps would be resolved. Yet many responsibility gaps extend beyond legal responsibility, often emerging in low-stakes AI contexts unlikely to be directly regulated by law. These contexts still involve real harms that necessitate allocation of blame or burden-sharing. How to equitably assign this responsibility in the moral realm remains unclear, preserving the persistence of responsibility gaps.

Consider the example of an email spam filter. Few consider spam filters pinnacle technology. Spam filters have been widespread for decades and represent decidedly low-stakes AI applications; hence they fall outside recent regulatory efforts such as the EU AI Act. Nevertheless, spam filters employ advanced machine-learning algorithms trained on vast data to classify emails, sometimes inaccurately and with harmful implications.

Spam Filter: You book a flight online, on the airline's website. After you see the booking confirmation on your screen, you close the tab on your browser and move on with your day. Naturally, you expect to receive an email with your booking confirmation. Yet, nothing appears in your inbox. Concerned that the payment did not go through, you repeat the same booking process, booking and paying for the same tickets again. This time you receive the confirmation email as expected. Later, you discover two identical charges from the airline on your bank account. You have paid twice for the same tickets. Browsing through your email spam folder, you find the original confirmation email by the airline, mistakenly classified as spam.

Who is responsible in a case like *Spam Filter*? The email provider, for a faulty filter? The airline, for poorly designed emails prone to misclassification? Or the user, who did not check spam before repurchasing? The answer is unclear, making this a paradigmatic AI-related responsibility gap: there is measurable harm caused but opaque distribution of blame and a lack of proper accountability. *Spam Filter* illustrates why regulation alone cannot fully resolve responsibility gaps, particularly in low-stakes AI applications. It also highlights a crucial distinction that is often overlooked in the literature on responsibility gaps: legal responsibility versus moral responsibility. Legal responsibility refers to duties enforceable by law, while moral responsibility arises from communal ethical norms that are typically unregulated.

The distinction between legal and moral responsibility demonstrates why *Spam Filter* is not a fringe case or an exception, but a rather commonplace example of everyday human-AI interaction. Moral responsibility and legal responsibility occupy spaces that intersect but are not identical. Therefore, regulation can only ever cover cases where there is a gap of legal—not moral—responsibility. To say that all responsibility gaps can be resolved by regulation is to equate moral responsibility with legal responsibility. Yet, most of our moral life is not regulated by laws, but by commonly accepted rules of engagement known as moral practice. Recent literature has explored how responsibility gaps may persist despite regulatory attempts to bridge them (see Coeckelbergh, 2020; Vallor & Vierkant, 2024). Regulation may resolve many high-stakes responsibility gaps. Yet low-stakes moral responsibility gaps common in everyday human-AI interactions are likely to persist.

3 EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)

3.1 THEORETICAL BACKGROUND AND CORE PRINCIPLES

Explainable Artificial Intelligence (XAI) has emerged as a response to the opacity of modern AI systems and the accompanying responsibility gaps. It encompasses a broad family of methods

aimed at generating explanations for how AI systems reach their outputs. These explanations seek to make AI decisions accessible and interpretable to users, thereby enabling transparency and, potentially, accountability. Over the past decade, XAI has gained substantial research interest, producing influential works ranging from foundational position pieces highlighting the necessity of interpretable models (Lipton, 2018), to comprehensive taxonomies classifying methods and concepts within XAI (Arrieta et al., 2020; Angelov et al., 2021), and recent surveys assessing current shortcomings and proposing future directions for the field (Ali et al., 2023; Haque et al., 2023). Fundamentally, XAI strives to “open the black box” inherent in complex machine learning models, shedding light on how and why specific outputs arise.

From both epistemological and ethical perspectives, XAI raises challenging questions. Good explanations should not only provide technical insights but should also foster understanding and enable human accountability. Regulatory frameworks and standards emphasize that quality explanations must be meaningful, accurate, and mindful of the system’s knowledge boundaries (Phillips et al., 2021). The National Institute of Standards and Technology (NIST), for example, articulates foundational principles of XAI including clear evidential support, audience-tailored comprehensibility, explanation accuracy, and respect for operational constraints.

A central tension in XAI concerns the trade-off between model accuracy and interpretability. While simple “white-box” models—such as decision trees or linear regressions—offer transparency, they typically underperform on complex tasks compared to opaque “black-box” models like deep neural networks (Ferlito et al., 2024). This trade-off complicates the attainment of fully explainable AI without compromising performance in critical domains such as healthcare or criminal justice.

Despite the promise of XAI methods, existing techniques face substantial limitations in fully closing the responsibility gap. Many explanations are post-hoc, posing challenges for broader ethical and normative assessments. The following section explores these epistemological limits and prepares the ground for Section 4, where I argue for the necessity for XAI to address two distinct but intertwined problems for truly bridging responsibility gaps. To begin, consider the proposed core principles for XAI according to the National Institute of Standards and Technology (NIST). According to NIST, good XAI should follow four *Foundational Principles*:

- *Explanation*: AI systems must provide evidence or reasoning supporting their outputs and internal processes. Explanations are not merely ancillary but are integral to understanding the system’s outputs.
- *Meaningful*: Explanations must be tailored to and comprehensible by the intended audience, which may range from technically qualified AI developers to non-expert end-users. This principle recognises that transparency requires accessibility.
- *Explanation Accuracy*: Explanations should accurately reflect the system’s decision-making process, avoiding misleading simplifications or distortions that undermine trust or accountability.
- *Knowledge Limits*: AI systems should operate strictly within their designed scope, including clear specifications of their confidence thresholds and limitations. This principle guards against overreliance on AI and inappropriate attribution of culpability (Phillips et al., 2021).

Together, the NIST Foundational Principles of XAI highlight that explainability is not merely a technical feature; it embodies a balancing act between efficacy, accessibility, and operational integrity. For example, a medical diagnostic AI system should not only justify its predictions with features demonstrating their importance but also contextualise these insights into clinically

actionable recommendations. These foundational principles provide a normative framework for the development and evaluation of XAI methods towards ensuring that explanations are not only technically sound but serve as a basis for human understanding and accountability.

3.2 XAI METHODS

The theoretical background behind Explainable Artificial Intelligence (XAI) is essential for understanding its scope, ambitions, and limitations. Central to this are the broad categories of techniques employed to make AI systems' decision-making processes more transparent. Here, I discuss three of the most prominent families of XAI methods. While this list is far from exhaustive, the three types of methods discussed here are an overall good representation of XAI in general. These are: *self-interpretable models*, *post-hoc explanation methods*, and *hybrid models*.

- **Self-Interpretable Models:** Self-interpretable models are models designed to be themselves explanations. Typically, these models include decision trees, logistic regression, and rule-based systems. Their key advantage lies in enabling direct inspection of their decision logic, providing global explanations of model behaviour (Alvarez-Melis & Jaakkola, 2018). For instance, decision trees partition the input space via hierarchical rules mapping features to outcomes, while logistic regression assigns explicit weights to features for transparent feature contribution assessment (Elton, 2020; Hulsén, 2023; Mahbooba et al., 2021). Rule-based systems simplify interpretation by encoding domain heuristics into explicit "if-then" statements (Van der Waa et al., 2021). Despite their interpretability strengths, self-interpretable models often underperform on complex, dynamic, or high-dimensional data compared to more opaque models. As a result, many researchers prefer models that are not themselves explanations, but rather models that are designed to generate explanations after training.
- **Post-Hoc Explanation Methods:** Post-hoc approaches analyse black-box models after training to produce interpretable explanation. Among these, LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) are the most notable. While post-hoc explanations exhibit considerable diversity, they all share the same objective: to analyse black-box models after their training has concluded (Guidotti et al., 2021; Retzlaff et al., 2024). More specifically, LIME generates local explanations by approximating model behaviour around specific inputs (Dieber & Kirrane, 2020; Salih et al., 2025). LIME achieves this using interpretable surrogate models (e.g., linear regressors) (Hung & Lee, 2024). That is, LIME essentially creates a synthetic dataset and trains a transparent model to explain the original opaque model. Conversely, SHAP employs coalitional game theory to assign feature importance scores, that is, the marginal contribution that each feature has to a prediction. These scores then inform the emerging post-hoc explanation (Heuillet et al., 2022; Salih et al., 2025). In contrast to LIME, which provides only local explanations, and self-interpretable models, which focus on global explanations, SHAP can provide both local and global explanations, thus ensuring consistency across different explanations (Agarwal et al., 2022; Leventi-Peetz & Weber, 2022; Slack et al., 2021). A third kind of post-hoc explanation method are counterfactual explanations (Dai et al., 2022; Stepin et al., 2021; Wachter et al., 2017). In contrast to LIME and SHAP, counterfactual explanations function by identifying minimal input perturbations that would alter a model's decision, thus offering actionable insights for users (Keane & Smyth, 2020; Dai et al., 2022).

LIME, SHAP, and counterfactual explanations are all model-agnostic methods, and they are widely applied in domains like healthcare and cybersecurity (Dai et al., 2022; Wachter et al., 2017).

- *Hybrid Models:* Hybrid models attempt to integrate interpretable components within complex architectures to balance accuracy and transparency (Tahir et al., 2024). Examples include neural networks augmented with decision trees or prototype-based classifiers, which offer interpretable rationales while maintaining competitive performance (Nauta et al., 2021). These hybrids are particularly beneficial in cybersecurity intrusion detection systems, where combining rule-based reasoning and deep learning enables anomaly detection alongside justifications for alerts (Kuppa & Le-Khac, 2020; Srivastava et al., 2022). However, hybrid methods can introduce computational and latency challenges that may hamper real-time deployment (Mohale & Obagbuwa, 2025).

3.3 LIMITATIONS OF POST-HOC EXPLANATIONS

Despite their widespread adoption and significant contributions to AI interpretability, post-hoc explanation methods face both practical and epistemological limitations. Post-hoc explanations are inherently retrospective; analysing models after training to explain past decisions and typically approximate, because they rely on surrogate models or simplifications rather than faithfully reconstructing the original system's exact computational pathways (Rudin & Radin, 2019; Guidotti et al., 2021). While retrospective explanations can be fully accurate for specific cases in principle, popular methods like LIME and SHAP generate approximations through local sampling, game-theoretic assumptions, or feature perturbations that trade fidelity for interpretability, often diverging from the true model logic (Alvarez-Melis & Jaakkola, 2018).

More specifically, two of the most prevalent post-hoc techniques, LIME and SHAP, suffer noteworthy shortcomings. LIME operates by sampling perturbed instances of the input to build a local surrogate interpretable model, but this sampling can lead to instability, producing different explanations on repeated runs with minimal input variance. SHAP, grounded in cooperative game theory, assumes independence among features (a condition often violated in real-world data) potentially biasing feature importance attributions (Bordt et al., 2022; Vale et al., 2022).

Counterfactual explanations, which aim to identify minimal changes to inputs required to alter an AI prediction, also face challenges. They often lack rigorous empirical validation and are vulnerable to generating unrealistic or uninformative scenarios without domain constraints (Keane et al., 2021). Moreover, all these methods risk instilling epistemic overconfidence, where users mistake post-hoc rationales for comprehensive understanding, thereby impairing accurate judgment and potentially exacerbating responsibility gaps when failures occur (Brand & Nannini, 2023; Bordt et al., 2022; Lima et al., 2022). In adversarial or high-stakes contexts (such as credit scoring, legal judgments, or medical diagnosis) the fragility and contestability of post-hoc explanations pose serious concerns. Explanations may be manipulated to justify biased or flawed decisions without genuine accountability. Thus, relying solely on post-hoc techniques can undermine trust and blur clear attribution of responsibility, aggravating the opacity they seek to mitigate (Bordt et al., 2022; Wachter et al., 2017).

Ultimately, the core issue arises from the “model-on-model” nature of post-hoc explanations: they are generated by a secondary model approximating the original system's behaviour. Any divergence or error in this approximation adds further uncertainty, questioning the epistemic reliability of explanations and limiting their sufficiency as a foundation for grounding moral or legal responsibility. These limitations underscore the persistent problem confronting XAI. While post-hoc explanations enhance transparency to some degree, they currently fall short of fully

bridging the responsibility gaps that arise from AI system opacity. In the next section I flesh out these problems in more detail.

4 TWO PROBLEMS FOR EXPLAINABLE AI

XAI represents an ambitious effort in human-AI alignment, where generated explanations serve as bridges between algorithmic processes and human comprehension. The aspirations of XAI researchers and supporting institutions—both public and private—revolve around alleviating key concerns tied to AI opacity and the black box problem. Foremost among these is the responsibility gap, which XAI appears uniquely positioned to address. Yet, as I argue in this section, XAI's effectiveness in bridging this gap depends on resolving two distinct yet interdependent challenges: an epistemological problem concerning the nature and adequacy of its explanations, and a moral attribution problem regarding the fair distribution of blame and praise among stakeholders. Recognising the distinct nature of these two problems is essential to fully grasp the space within which XAI is meant to provide solutions. Unfortunately, current XAI techniques fall short of fully tackling both the epistemological problem and the moral attribution problem, leaving the responsibility gap at least partly unresolved.

4.1 THE EPISTEMOLOGICAL PROBLEM: FROM EXPLANATION TO UNDERSTANDING

XAI aims to make AI decision-making interpretable by generating explanations of outputs. However, these explanations often fail to meet the epistemic standards required for grounding moral or legal responsibility, as surveyed in Section 3. Two primary types dominate the literature: causal explanations, which identify input features that causally produce specific outputs (Baron, 2023; Beckers, 2022; Chou et al., 2022), and reason explanations, which highlight features serving as justifying reasons for outcomes (D. Baum et al., 2023; Robbins, 2019). Both are predominantly post-hoc, analysing trained models retrospectively, and inherit characteristic limitations that undermine their adequacy for responsibility attribution (Arrieta et al., 2020; Guidotti et al., 2021). Resolving this epistemological problem requires overcoming three interrelated issues.

First, post-hoc explanations may be responsible for creating a false sense of understanding. Instead of solving the epistemological problem, such explanations may instead make the problem more pressing than it was before the explanation was generated. Leading XAI methods like feature attribution (SHAP/LIME) and counterfactuals generate post-hoc rationalizations approximating model behaviour, rather than revealing intrinsic decision processes (Rudin & Radin, 2019; Guidotti et al., 2021). Consider a medical diagnostic AI flagging biomarkers as pivotal for cancer detection: this surrogate insight obscures whether the model's logic mirrors clinical reasoning or exploits spurious correlations (Lima et al., 2022; Taylor, 2024). Such approximations foster “epistemic overconfidence” or a false sense of understanding, where users may overlook both the model's and their own biases or limits, in favour of a relatively simple and comprehensible explanation of the model's reasoning (Brand & Nannini, 2023; Arrieta et al., 2020). This may lead to inflated or misplaced trust in incomplete or meaningfully flawed explanations, exacerbating responsibility gaps whenever a model fails (Ferlito et al., 2024; Taylor, 2024).

Second, post-hoc explanations often involve a trade-off between *accuracy* and *interpretability*. High-performance AI systems, particularly deep learning models, often sacrifice transparency for accuracy, facing common “black-box AI” objections. Conversely, “white-box” alternatives (e.g., models using linear regression) are largely interpretable, yet they often underperform compared to opaque “black-box” models in complex tasks like natural language processing (Rudin & Radin, 2019) and medical imaging analysis (Ferlito et al., 2024). Hybrid compromises

exist, yet regulatory explainability mandates risk excluding superior opaque models from high-stakes domains, entrenching existing responsibility gaps where opacity persists by design (Doshi-Velez & Kim, 2017).

Third, post-hoc explanations are often both too technical and not sufficiently grounded in the context in which they operate, thus running the risk of becoming largely irrelevant. XAI typically prioritises technical metrics (e.g., feature importance scores) over explanatory or moral saliency. For example, an AI used in credit decisioning might cite income level as a decisive factor for approving or denying someone a loan. Yet this often overlooks systemic biases present in training data or historic social inequalities, making explanations normatively incontestable (Selbst & Barocas, 2018; Lima et al., 2022; Brand & Nannini, 2023). Robust explanations must bridge to ethical principles like fairness or justice; current frameworks rarely achieve this, leaving users unable to evaluate decisions morally and perpetuating attribution ambiguity (Yadav, 2024). These epistemological shortcomings suggest that XAI explanations, while useful, are insufficient for bridging responsibility gaps unless they are redesigned to prioritise actionable and morally contextualised insights. While some recent work proposes that certain “reason-giving” XAI techniques (primarily explaining decisions through normatively structured rationales) might enable responsibility attribution (K. Baum et al., 2022), such approaches remain exceptional amid predominantly technical post-hoc methods. Their success depends on overcoming the contextual limitations documented above, suggesting that broader systemic redesign remains necessary.

4.2 THE MORAL ATTRIBUTION PROBLEM: DISTRIBUTING BLAME AND PRAISE

XAI is a project aimed at making opaque systems more transparent to human users. However, when the broader goal is to bridge gaps of responsibility, system transparency alone is not sufficient. Apart from making models more transparent, XAI should also enable humans to find the best way to distribute burdens among affected parties in scenarios where AI is involved in decision making. In other words, a good XAI explanation is one that assists us in attributing blame and praise to the appropriate targets. Such an explanation should be able to tackle the moral attribution problem. Even if XAI provided epistemologically robust explanations, if the moral attribution problem is not resolved, responsibility gaps would persist. Currently, the moral attribution problem remains unresolved. There are three main factors at play here:

- *The Problem of Many Hands*: AI systems operate within complex socio-technical networks involving developers, data scientists, users, regulators, and institutions, each contributing causally yet opaquely to outcomes (Slota et al., 2023). This is what is commonly known as the Problem of Many Hands (PMH) (for the origins of the problem, see Thompson, 1980). When harms arise, responsibility diffuses across these “many hands”, obscuring clear attribution of responsibility (Van de Poel, 2015). For instance, a clinical AI misdiagnosis might trace to flawed training data (developer oversight), inadequate regulatory validation (agency failure), biased deployment protocols (institutional choices), or clinician overreliance (user error) (Berber & Mijić, 2024; Coeckelbergh, 2020). XAI explanations illuminate the model’s specific contributions (e.g., feature weights) but they fail to disentangle upstream causal chains or apportion collective versus individual culpability, particularly when harms stem from systemic interactions rather than isolated negligence (Kiener, 2025; Van de Poel, 2015). Philosophers distinguish PMH from simple causal complexity by its normative dimension: not just “who caused it,” but “who ought to bear remedial burdens” (Miller, 2007). In AI contexts, this demands tracing distributed agency across lifecycles, yet current frameworks remain individualist, ill-suited to collective AI governance (Slota et

al., 2023). Recent proposals advocate “responsibility-sharing matrices” or role-specific duties, but integration with XAI remains underdeveloped (Hindriks & Veluwenkamp, 2023). Thus, PMH persists as a structural barrier to attribution, beyond explanatory transparency alone.

- The Tension Between Control and Knowledge:* Moral responsibility traditionally requires both control over one’s actions and knowledge of the action’s potential moral consequences (Fischer & Ravizza, 1998; Vargas, 2013). XAI enhances epistemic knowledge by clarifying decision pathways to inform users. But it often leaves agents without meaningful control over AI outputs or underlying algorithms. For example, a radiologist using a diagnostic AI may fully understand its explanations (e.g., “80% tumour probability due to lesion size”), yet face practical barriers (institutional protocols, time pressures, automation bias, or deference to algorithmic authority) that compromise her ability to go against the AI recommendation, even when that would be clinically warranted (Coeckelbergh, 2020; Taylor, 2024). While she retains formal authority to override the AI, genuine decision control requires a realistic capacity to act against its suggestions despite these pressures. This creates an accountability mismatch: end-users bear liability for decisions they oversee but cannot alter, while developers evade responsibility by invoking model autonomy or complexity (Danaher, 2016). In legal terms, this echoes problems of “automation bias,” where understood-but-uncontrollable AI shifts moral burdens unpredictably. Philosophical accounts distinguish direct control (e.g., altering code) from indirect control or guidance (e.g., input validation), but AI ecosystems complicate both. Users possess neither, trapped in a “responsibility without power” dilemma that backward-looking blame models exacerbate (Nyholm, 2023). Recent analyses frame this as a “control gap” complementing epistemic gaps, urging hybrid responsibility frameworks that recalibrate expectations across the AI lifecycle (Hindriks & Veluwenkamp 2023; Kiener, 2022). Resolving this tension demands not just better explanations, but redefined control paradigms.
- The Limits of Backward-Looking Responsibility:* Dominant accounts of moral responsibility in contemporary philosophy are predominantly backward-looking, focusing on retrospective blame or praise after harm has occurred (Strawson, 1962; Fischer & Ravizza, 1998). Conversely, forward-looking accounts, sometimes termed “remedial responsibility” in political theory (Miller, 2007), emphasise prospective duties to prevent future harms or rectify ongoing issues, often irrespective of prior fault. However, AI systems demand primarily proactive, forward-looking mechanisms to anticipate, mitigate, and allocate risks before deployment or failure. Backward-looking frameworks, while illuminating retrospective accountability (e.g., blaming developers post-harm), struggle to prescribe preventive burden-sharing in dynamic AI ecosystems, where harms often emerge from emergent behaviours rather than from isolated faults. For instance, ensuring safe autonomous vehicles requires ongoing remedial duties among manufacturers, regulators, and users. Such duties are not easily derived from blame-centric models (Hindriks & Veluwenkamp, 2023). Philosophers have long discussed forward-looking responsibility, yet its integration into AI ethics remains nascent and fragmented. While backward-looking insights can inform forward ones (e.g., via Strawsonian “reactive attitudes” adapted prospectively), the challenge lies in operationalising hybrid models tailored to the nuances of AI, like distributed agency and scalability (Coeckelbergh, 2020; Kiener, 2022). Recent work highlights “control gaps” as a complement to responsibility gaps, underscoring the need for novel frameworks (Hindriks & Veluwenkamp, 2023).

This underscores the imperative for forward-looking accounts attuned to AI, though developing them lies beyond the scope of the paper.

5 TWO PROPOSALS AND A WORRY

The aim of this next section is twofold. First, to present two proposals to overcome the current predicament of XAI. The first is a possible solution to the epistemological and moral attribution problems presented in Section 4. It acknowledges the two problems as separate but interdependent and suggests specific reforms to tackle them in tandem. The second is a recent alternative approach to AI interpretability which may be better positioned than traditional XAI to bridge responsibility gaps. Finally, I introduce and briefly discuss a set of worries surrounding the XAI research program, its ambitions and limitations.

5.1 A DUAL SOLUTION TO THE TWO PROBLEMS

Current XAI falls short of bridging the responsibility gap. To do that, XAI needs to confront both the epistemological limitations of its explanations and the complexities of attributing moral responsibility in the real world. Bridging the responsibility gap demands addressing both problems in tandem, with a set of reforms to tackle both the epistemological and the moral attribution problem. Together, these reforms present at least a partial dual solution to the two problems of XAI. They consist of *epistemological reforms* and *moral attribution reforms*.

5.1.1 EPISTEMOLOGICAL REFORMS

To address the epistemological problem, XAI should make the shift from generating technical explanations to providing normative justifications. This may involve a number of interventions into current XAI. I briefly discuss four potential interventions below:

- **Contextualizing explanations within domain-specific ethical frameworks**, for example “medical beneficence” or “financial fairness”. Achieving this requires the parallel adoption of two approaches: (a) interface design where explanations dynamically map technical outputs to ethical concepts—for instance, a medical AI not just listing “biomarker weights” but anchoring them to beneficence (“this feature elevates risk, prioritising harm prevention per HIPAA norms”); and (b) user training modules integrated into deployment, teaching stakeholders to spot ethical misalignments (e.g., fairness violations in lending) (Doshi-Velez & Kim, 2017; Floridi et al., 2018). These build on existing work and extend it by embedding normative templates directly in model pipelines.
- **Prioritizing interpretable models without sacrificing performance**, wherever that is possible. For instance, adopting hybrid systems that combine opaque and transparent components. At the same time, opting for models that enable “explanation-preserving distillation,” that is distilling black-box accuracy into glass-box proxies while auditing its fidelity to the original model (Tahir et al., 2024).
- **Enhancing user literacy to critically evaluate explanations**, thus ensuring stakeholders can discern between *approximate explanations* and *reliable reasoning*. One way to achieve this is by deploying standardised “explanation audits” (e.g., confidence intervals on SHAP values) alongside various types of interactive tutorials.
- **Contextualizing explanations according to different stakeholders**. Tailored explanations can provide a desirable degree of granularity for each group. For instance, developers may receive circuit-level diagnostics, end-users might get

natural-language ethical summaries, and regulators could receive compliance traces, all corresponding to their respective epistemic needs and roles (Miller, 2007).

5.1.2 MORAL ATTRIBUTION REFORMS

To address the moral attribution problem, the following four reforms are proposed, operating at distinct but interdependent levels. The first two concern moral practice: reframing the ontological assumptions underlying responsibility attribution and cultivating the practical capacities required for responsible AI use. The third moves toward institutional redesign, proposing context-sensitive standards for distributing responsibility across different AI applications. The fourth touches on possible legal reform, exploring mechanisms for codifying stakeholder obligations in ways that could carry either normative or legal force. Together, these reforms do not constitute a complete solution to the moral attribution problem, but they represent a substantive step toward the kind of systemic, forward-looking accountability that bridging responsibility gaps demands.

A first moral attribution reform concerns the ontological assumptions underlying how we distribute responsibility in AI contexts. Dominant Western frameworks of moral responsibility are built around the individual as the primary unit of agency: responsibility is assigned to discrete persons who acted with sufficient control and knowledge. This individualist ontology is structurally mismatched with the distributed, socio-technical character of AI systems, where agency is diffused across developers, datasets, institutions, users, and potentially the systems themselves. No single individual occupies the position that individualist responsibility models presuppose, which is precisely why the Problem of Many Hands proves so intractable within these frameworks.

The African philosophical tradition of Ubuntu offers a structurally better fit. Ubuntu's foundational principle—*umuntu ngumuntu ngabantu*, roughly “I am because we are”—suggests that personhood and agency are constitutively relational rather than atomistically individual. On this view, the responsible subject is not a discrete individual but a network of interdependent agents whose identities and obligations are mutually constituted. This is not merely a cultural alternative to Western individualism but an ontological reframing of who bears responsibility and how burdens should be distributed. Crucially, Ubuntu emphasises communal prevention over retrospective personal blame: the question is not only who caused the harm, but how the community of agents failed collectively to prevent it. This maps directly onto the distributed agency characteristic of AI systems and offers a principled basis for the kind of systemic, forward-looking accountability that backward-looking individualist models cannot provide. It is worth distinguishing this proposal from existing appeals to collective responsibility in the AI ethics literature, such as those concerning autonomous military systems or responsibility loci in human-robot collaboration (Conradie, 2023; Nyholm, 2018). Those accounts extend Western collective responsibility frameworks to AI contexts without challenging their underlying individualist ontology. The Ubuntu proposal goes further: it contests the ontological baseline itself, arguing that genuinely bridging responsibility gaps requires rethinking the nature of the responsible subject, not merely redistributing blame among a larger set of individuals (Metz, 2011; Ferlito et al., 2024).

A second reform concerns the cultivation of the capacities required for responsible AI use. Even if ontological reframing shifts our understanding of who the responsible subject is, this is insufficient unless the agents involved actually possess the practical capabilities to act responsibly within AI-assisted contexts. This points to a foundational role for what Victoria McGeer calls the “agency cultivation” model of responsibility: rather than merely assigning blame after the fact, responsibility practices should actively foster the capacities for foresight, judgment, and accountability across all stakeholders in the AI lifecycle (McGeer, 2015; Vargas, 2013).

The Aristotelian concept of *phronesis* (typically translated as practical wisdom) provides the appropriate philosophical grounding here. *Phronesis* is not a set of rules to be followed, but a stable disposition to deliberate well about how to act in particular circumstances, developed through habituation and experience rather than abstract instruction (Aristotle, *Nicomachean Ethics*, VI). Responsible engagement with AI systems requires precisely this kind of cultivated judgment: not merely knowing that a medical AI flags certain biomarkers, or that a hiring algorithm weighs certain features, but developing the practical wisdom to evaluate those outputs critically, identify when they should be overridden, and act accordingly despite institutional pressures or automation bias. This is a capacity that must be developed over time and across contexts. AI-related *phronesis* cannot be reduced to a one-off training module or a compliance checklist. Concretely, agency cultivation might take the form of iterative feedback mechanisms in which developers simulate edge cases and stress-test their own models, users receive structured opportunities to practice critical evaluation of AI outputs, and regulators develop domain-specific expertise rather than relying solely on technical audits. These mechanisms operationalise *phronesis* as an institutional practice rather than merely an individual virtue, making it available as a precondition for the other reforms proposed here to function effectively.

A third reform concerns the development of context-specific responsibility standards for common AI applications where responsibility gaps are likely to arise. Rather than applying a single uniform standard across all AI contexts, we propose that the appropriate distribution of responsibility should be determined by three criteria: the stakes involved, the degree of user control, and the foreseeability of harm. Varying combinations of these criteria point toward different responsibility standards, of which two are particularly relevant here.

The first is a *strict liability* standard, applicable where stakes are high, user control is low, and harm is foreseeable. In such cases, burden-shifting away from the injured party is morally warranted, and responsibility falls primarily on developers and deploying institutions. A medical AI that misdiagnoses a patient is a paradigmatic case: the harm is serious, the user (typically a clinician) has limited ability to independently verify the system's internal reasoning, and the risk of harm from diagnostic error is entirely foreseeable. Strict liability assigns responsibility to those best positioned to prevent it, namely the developers and institutions that designed, validated, and deployed the system.

The second is a *user-beware* standard, applicable where stakes are modest, user control is meaningful, and the system's limitations are either known or reasonably knowable. Here, the expectation of independent judgment on the part of the user is morally defensible, and responsibility for harms arising from uncritical reliance falls at least partly on the user. Consider a navigation AI that routes a user through an unsafe neighbourhood: as Stamatiou and Karakonstantis (2024) argue, the appropriate attitude toward such systems is one of calibrated reliance rather than trust. A navigation app makes no claim to guarantee safe passage; its scope is well-defined and widely understood. A user who extends to it quasi-trust, deferring entirely to its recommendations without exercising independent situational judgment, has over-relied on the system in a way that is partly their own responsibility. The *Spam Filter* case discussed in Section 2 illustrates the same standard at the lower end of the stakes spectrum: the harm is real but modest, and a reasonable degree of user vigilance (checking the spam folder for expected correspondence) is both possible and appropriate.

Between these two poles sits a range of intermediate cases where neither standard applies cleanly. A hiring algorithm that rejects a candidate occupies precisely this middle ground: the stakes are significant for the individual affected, user control varies considerably depending on whether a human reviewer is meaningfully involved, and the foreseeability of discriminatory outcomes is well-documented but unevenly acknowledged across institutions. Such cases call for hybrid standards that combine elements of institutional accountability with user and

operator obligations, and they underscore the need for ongoing normative work to develop more fine-grained criteria as AI applications proliferate.

A fourth reform concerns the codification of stakeholder-specific obligations through what the literature refers to as “responsibility contracts” (Van de Poel, 2015). The diffusion of responsibility across the many hands involved in AI development and deployment is partly a structural problem, but it is also partly a problem of normative ambiguity: stakeholders often lack clear, acknowledged obligations that specify what responsible conduct requires of them at each stage of the AI lifecycle. Responsibility contracts address this by making obligations explicit, turning compliance from a one-sided regulatory requirement into a shared normative commitment among all parties involved.

Such contracts could take at least two forms, and the choice between them carries significant practical implications. The first is a soft governance instrument: a voluntary, sector-specific code of conduct in which developers, deploying institutions, and users acknowledge their respective obligations. For instance, developers ensuring model robustness and documentation, institutions validating outputs against domain-specific standards, and users committing to critical engagement rather than uncritical deference. Professional codes of conduct in medicine or law offer a useful analogue: they are not legally binding in every instance, but they carry normative force within their respective communities and provide a basis for professional accountability. The second is a legally binding instrument, closer in character to data processing agreements under the GDPR, in which obligations are formally specified, enforceable, and subject to regulatory oversight. This model offers stronger accountability guarantees but requires more developed institutional infrastructure and may be better suited to high-stakes domains where the strict liability standard discussed above already applies.

Which of these forms is ultimately more appropriate is a question that cannot be resolved in the abstract; it will depend on the domain, the maturity of relevant regulatory frameworks, and the degree of normative consensus among stakeholders. What matters for present purposes is that responsibility contracts of either kind represent a concrete mechanism for operationalizing the other reforms proposed here: translating the ontological reframing of Ubuntu, the cultivated judgment of phronesis, and the context-specific standards into acknowledged, actionable obligations that can be monitored, revised, and held accountable.

To be sure, providing an exhaustive list of reforms or proposals is beyond the scope of the paper. Still, according to the dual solution sketched out here, technical transparency must be paired with ethical contextualization. At the same time, blame attribution requires moving beyond individual culpability to systemic accountability. Future research should prioritise XAI methods that align explanations with normative reasoning and governance models that distribute responsibility across stakeholders equitably. Only then can AI systems operate as morally integrated tools rather than sources of ethical ambiguity.

5.2 MECHANISTIC INTERPRETABILITY

Mechanistic interpretability (MI) refers to a family of methods that aim to explain how neural networks work by uncovering their underlying causal mechanisms (Williams et al., 2025). This often involves reverse-engineering specific circuits associated with certain operations, like object detection in vision models (Templeton et al., 2024). While the MI research program shares many of the goals of XAI, some have suggested that it also departs from it in important ways (see Chalmers, 2025). For instance, MI prioritises intrinsic causal maps over post-hoc surrogate explanations dominant in XAI. MI has recently been proposed as a potential technical solution to responsibility gaps by making it easier to trace how a system arrived at a harmful or contested output (Chalmers, 2025; Williams et al., 2025; Templeton et al., 2024).

At the same time, the discussion of MI as a viable alternative to XAI raises several deeper philosophical questions: Can MI improve the capacity of regulators or users to attribute responsibility in institutional settings? Is MI also able to support informal moral practices of blame and praise in everyday life? How does MI compare to other forms of AI explainability when it comes to ethical adequacy? While these questions remain relatively under-explored and largely beyond the scope of this paper, philosophical engagement with them could help shape how we interpret MI's normative potential in both high-stakes and low-stakes domains.

5.3 A WORRY: IS XAI WHITEWASHING THE TECH INDUSTRY?

To all appearances, the goals of XAI are noble: to develop, test and apply methods that make opaque models interpretable. However, one can plausibly worry that the push for explainability and interpretability could be used as a convenient facade for the AI industry's reluctance to accept responsibility for their own creations. There are two ways to flesh out this worry. First, let us accept that black box AI is indeed as much a "black box" as many claim. Even if that is true, black boxes are not spontaneously generated. They are the result of technological advancement that is both intentional and extremely sophisticated. So, if indeed AI is as opaque as some claim, and if that is causing problems like gaps of responsibility, why do we keep building it that way? That is not to say that we should retreat to some kind of "epistemological luddism" as Langdon Winner has claimed (Winner, 1978), but at a minimum, it suggests we should rethink how we hold the providers of opaque AI accountable.

But one can also ask whether the black box nature of AI is itself exaggerated. Indeed, models that use deep learning tend to be more opaque than others, but the narrative of opacity around AI could be overblown. After all, most technology is opaque in at least some degree, even to its own designers. Car manufacturers have recognised this for decades. That is why cars are vigorously tested in a variety of different scenarios, including crash tests using crash dummies, to approximate what would happen to human passengers during a crash. According to how well it performs in these tests, each model is assigned unique safety scores, which affect their pricing, commercial success, and ultimately, the manufacturer's reputation. But for those in the AI industry who wish it to remain largely unregulated, an inflated sense of opacity has several benefits. It allows AI providers (or at least some of them) to design, develop, and deploy models without worrying too much about their potentially harmful implications (Zuboff, 2019). The narrative on responsibility gaps provides the perfect cover-up for those who are not interested in being held accountable for the defects of their products. At the same time, the rest of us are complicit in allowing them to turn something that should be a problem for the technology, into a problem for everyone.

6 CONCLUSION

A thread that runs through the entire paper, and that deserves to be made explicit in closing, is the distinction between legal and moral responsibility introduced in Section 2. Regulatory efforts such as the EU AI Act represent significant progress in addressing responsibility gaps in high-stakes AI domains. Yet, as the Spam Filter case illustrates, a large and consequential class of responsibility gaps arises in precisely the contexts that regulation does not and cannot reach: everyday, low-stakes human-AI interactions where real harms occur but no legal framework assigns accountability. It is in this space of moral responsibility gaps that XAI's promise is most acutely felt and most acutely frustrated. The epistemological and moral attribution reforms proposed in Section 5 are aimed specifically at this space. They do not compete with legal regulation but complement it, addressing the moral residue that regulation inevitably leaves behind. Without advances on both fronts—legal frameworks for high-stakes domains and reformed moral practices for the everyday encounters that fall outside them—the responsibility

gap will persist, not as an abstract philosophical puzzle, but as a lived feature of our interactions with AI systems.

In this paper I have argued that the responsibility gap in explainable artificial intelligence (XAI) remains a challenge, undermining hopes that explainability alone can restore human accountability in AI-driven decisions. While XAI methods offer significant technical advances (see Section 3), they fail to deliver epistemically robust explanations capable of grounding moral responsibility because of their limitations (Section 4.1). Compounding this, even perfect explanations face significant challenges in assisting with distributing blame and praise, thereby creating a moral attribution problem for XAI (Section 4.2).

The dual reforms proposed in Section 5.1 represent interdependent steps forward, but they require rethinking both XAI pipelines and common moral practice. Mechanistic interpretability (MI) emerges as a promising complementary project, though its scalability and normative fit require philosophical scrutiny. Yet these remain proposals, not solutions; their success hinges on empirical validation and institutional adoption.

Ultimately, without significant advances, XAI risks perpetuating the harms of opacity, from spam filter mishaps to clinical misdiagnoses. At the same time, industry narratives tend to deflect accountability. Bridging responsibility gaps demands collective resolve to prioritise interpretable-by-design systems and fair sharing of potential burdens. This requires interdisciplinary efforts, like merging XAI engineering, epistemology, and ethics. Only then can XAI fill its goal of providing explanations that enable human accountability and agency.

Data Access Statement

N/A

Contributor Statement

Filippos Stamatiou: Conceptualization, Methodology, Writing – original draft, Writing – review & editing.

Use of AI

During the preparation of this work, the author(s) used a Claude (Anthropic) for assistance with final draft copy editing. After using this tool, the author reviewed, edited, made the content their own and validated the outcome as needed, and takes full responsibility for the content of the publication.

Funding Statement

N/A

Acknowledgments

An earlier version of this paper was presented at the workshop ‘Ethics and Normativity of Explainable AI: Issues in XAI #7’ (Paderborn University, 16 May 2024). The author thanks the organisers and participants for their feedback.

Conflict of Interest Statement

There is no conflict of interest.

References

- Agarwal, C., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., & Lakkaraju, H. (2022). “OpenXAI: Towards a Transparent Evaluation of Model Explanations.” *Advances in Neural Information Processing Systems*, 35, 15784–15799. <https://doi.org/10.52202/068431-1148>

- Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99, 101805. <https://doi.org/10.1016/j.inffus.2023.101805>
- Alvarez-Melis, D., & Jaakkola, T. S. (2018). Towards robust interpretability with self-explaining neural networks. *Advances in Neural Information Processing Systems*, 31. <https://doi.org/10.48550/ARXIV.1806.07538>
- Angelov, P. P., Soares, E. A., R. Jiang, Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: An analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5), e1424. <https://doi.org/10.1002/widm.1424>
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Benetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Baron, S. (2023). Explainable AI and causal understanding: Counterfactual approaches considered. *Minds and Machines*, 33(2), 347–77. <https://doi.org/10.1007/s11023-023-09637-x>
- Bathae, Y. (2018). The artificial intelligence black box and the failure of intent and causation. *Harvard Journal of Law & Technology*, 31(2), 889.
- Baum, D., Baum, K., Gros, T. P., & Wolf, V. (2023). XAI requirements in smart production processes: A case study. Longo (Ed.), *Explainable Artificial Intelligence* (Vol. 1901, pp. 3–24). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-44064-9_1
- Baum, K., Mantel, S., Schmidt, E., & Speith, T. (2022). From responsibility to reason-giving explainable artificial intelligence. *Philosophy & Technology*, 35, 12. <https://doi.org/10.1007/s13347-022-00510-w>
- Beckers, S. (2022). Causal explanations and XAI. Proceedings of the First *Conference on Causal Learning and Reasoning in Proceedings of Machine Learning Research*, 177, 90–109. <https://proceedings.mlr.press/v177/beckers22a.html>
- Berber, A., & Mijić, J. (2024). Understanding moral responsibility in automated decision-making: Responsibility gaps and strategies to address them. *Theoria*, 67(3), 177–92. <https://doi.org/10.2298/THEO2403177B>
- Bordt, S., Finck, M., Raidl, E., & Von Luxburg, U. (2022). Post-Hoc explanations fail to achieve their purpose in adversarial contexts. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 891–905. <https://doi.org/10.1145/3531146.3533153>
- Brand, J. L. M., & Nannini, L. (2023). Does Explainable AI Have Moral Value? arXiv. <https://doi.org/10.48550/arXiv.2311.14687>

- Buckley, R. P., Zetzsche, D. A., Arner, D. W., & Tang, B. W. (2021). Regulating Artificial Intelligence in Finance: Putting the Human in the Loop. *Sydney Law Review*, 43(1), 43–81.
- Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 2053951715622512. <https://doi.org/10.1177/2053951715622512>
- Chalmers, D. J. (2025). *Propositional Interpretability in Artificial Intelligence*. arXiv. <https://doi.org/10.48550/arXiv.2501.15740>
- Chou, Y.-L., Moreira, C., Bruza, P., Ouyang, C., & Jorge, J. (2022). Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion*, 81, 59–83. <https://doi.org/10.1016/j.inffus.2021.11.003>
- Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics*, 26(4), 2051–68. <https://doi.org/10.1007/s11948-019-00146-8>
- Conradie, N. H. (2023). Autonomous military systems: Collective responsibility and distributed burdens. *Ethics and Information Technology*, 25(1), 20. <https://doi.org/10.1007/s10676-023-09696-9>
- Dai, X., Keane, M. T., Shalloo, L., Ruelle, E., & Byrne, R. M. J. (2022). Counterfactual explanations for prediction and diagnosis in XAI. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 215–226. <https://doi.org/10.1145/3514094.3534144>
- Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology*, 18(4): 299–309. <https://doi.org/10.1007/s10676-016-9403-3>
- Danaher, J. (2022). Tragic choices and the virtue of techno-responsibility gaps. *Philosophy & Technology*, 35(2), 26. <https://doi.org/10.1007/s13347-022-00519-1>
- Dieber, J., & Kirrane, S. (2020). *Why model why? Assessing the strengths and limitations of lime*. arXiv. <https://doi.org/10.48550/arXiv.2012.00093>
- Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv. <https://doi.org/10.48550/arXiv.1702.08608>
- Elton, D. C. (2020). Self-Explaining AI as an Alternative to Interpretable AI. In B. Goertzel, A. I. Panov, A. Potapov, & R. Yampolskiy (Eds.), *Artificial General Intelligence* (Vol. 12177, pp. 95–106). Springer International Publishing. https://doi.org/10.1007/978-3-030-52152-3_10
- Ferlito, B., Segers, S., De Proost, M., & Mertes, H. (2024). Responsibility gap(s) due to the introduction of AI in healthcare: An ubuntu-inspired approach. *Science and Engineering Ethics*, 30(4), 34. <https://doi.org/10.1007/s11948-024-00501-4>
- Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge University Press.
- Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a

- good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>
- Guidotti, R., Monreale, A., Pedreschi, D., & Giannotti, F. (2021). Principles of explainable artificial intelligence. In M. Sayed-Mouchaweh (Ed.), *Explainable AI Within the Digital Transformation and Cyber Physical Systems* (pp. 9–31). Springer International Publishing.
- Haque, A. B., Islam, A. K. M. N., & Mikalef, P. (2023). Explainable artificial intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technological Forecasting and Social Change*, 186, 122120. <https://doi.org/10.1016/j.techfore.2022.122120>
- Heuillet, A., Couthouis, F., & Díaz-Rodríguez, N. (2022). Collective explainable AI: Explaining cooperative strategies and agent contribution in multiagent reinforcement learning with Shapley values. *IEEE Computational Intelligence Magazine*, 17(1), 59–71. <https://doi.org/10.1109/MCI.2021.3129959>
- Hindriks, F., & Veluwenkamp, H. (2023). The risks of autonomous machines: From responsibility gaps to control gaps. *Synthese*, 201(1), 21. <https://doi.org/10.1007/s11229-022-04001-5>
- Hulsen, T. (2023). Explainable artificial intelligence (XAI): Concepts and challenges in healthcare. *AI*, 4(3), 652–666. <https://doi.org/10.3390/ai4030034>
- Hung, Y. H., & Lee, C. Y. (2024). BMB-LIME: LIME with modeling local nonlinearity and uncertainty in explainability. *Knowledge-Based Systems*, 294, 111732. <https://doi.org/10.1016/j.knosys.2024.111732>
- Keane, M. T., Kenny, E. M., Delaney, E., & Smyth, B. (2021). *If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques*. arXiv. <https://doi.org/10.48550/arXiv.2103.01035>
- Keane, M. T., & Smyth, B. (2020). Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI (XAI). In I. Watson & R. Weber (Eds.), *Case-Based Reasoning Research and Development* (Vol. 12311, pp. 163–178). Springer International Publishing. https://doi.org/10.1007/978-3-030-58342-2_11
- Kiener, M. (2022). Can we bridge AI's responsibility gap at will? *Ethical Theory and Moral Practice*, 25(4), 575–93. <https://doi.org/10.1007/s10677-022-10313-9>
- Kiener, M. (2025). AI and responsibility: No gap, but abundance. *Journal of Applied Philosophy*, 42(1), 357–74. <https://doi.org/10.1111/japp.12765>
- Königs, P. (2022). Artificial intelligence and responsibility gaps: What is the problem? *Ethics and Information Technology*, 24(3), 36. <https://doi.org/10.1007/s10676-022-09643-0>
- Kuppa, A., & Le-Khac, N. A. (2020). Black box attacks on explainable artificial intelligence(XAI) methods in cyber security. *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8. <https://doi.org/10.1109/IJCNN48605.2020.9206780>

- Lang, B. H., Nyholm, S., & Blumenthal-Barby, J. (2023). Responsibility gaps and black box healthcare AI: Shared responsabilization as a solution. *Digital Society*, 2(3), 52. <https://doi.org/10.1007/s44206-023-00073-z>
- Leventi-Peetz, A.-M., & Weber, K. (2022). Rashomon Effect and consistency in explainable artificial intelligence (XAI). In K. Arai (Ed.), *Proceedings of the Future Technologies Conference (FTC) 2022, Volume 1* (Vol. 559, pp. 796–808). Springer International Publishing. https://doi.org/10.1007/978-3-031-18461-1_52
- Lima, G., Grgić-Hlača, N., Jeong, J. K., & Cha, M. (2022). The conflict between explainable and accountable decision-making algorithms. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2103–13. <https://doi.org/10.1145/3531146.3534628>
- Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3), 31–57. <https://doi.org/10.1145/3236386.3241340>
- Mahbooba, B., Timilsina, M., Sahal, R., & Serrano, M. (2021). Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. *Complexity*, 2021(1), 6634811. <https://doi.org/10.1155/2021/6634811>
- Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6, 175–83. <https://doi.org/10.1007/s10676-004-3422-1>
- McGeer, V. (2015). Building a better theory of responsibility. *Philosophical Studies*, 172, 2635–2649. <https://doi.org/10.1007/s11098-015-0478-1>
- Metz, T. (2011). Ubuntu as a moral theory and human rights in South Africa. *African Human Rights Law Journal*, 11(2), 532–59.
- Miller, D. (2007). The responsibility to protect human rights. In *Legitimacy, Justice and Public International Law*, 48.
- MirzaeiGhazi, S., & Stenseke, J. (2025). Responsibility before freedom: Closing the Responsibility gaps for autonomous machines. *AI and Ethics*, 5(2), 1601–1613. <https://doi.org/10.1007/s43681-024-00503-9>
- Mohale, V. Z., & Obagbuwa, I. C. (2025). A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. *Frontiers in Artificial Intelligence*, 8, 1526221. <https://doi.org/10.3389/frai.2025.1526221>
- Munch, L., Mainz, J., & Bjerring, J. C. (2023). The value of responsibility gaps in algorithmic decision-making. *Ethics and Information Technology*, 25(1), 21. <https://doi.org/10.1007/s10676-023-09699-6>

- Nauta, M., Van Bree, R., & Seifert, C. (2021). Neural prototype trees for interpretable fine-grained image recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14928–14938. <https://doi.org/10.1109/CVPR46437.2021.01469>
- Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Science and Engineering Ethics*, 24(4), 1201–1219. <https://doi.org/10.1007/s11948-017-9943-x>
- Nyholm, S. (2023). Responsibility gaps, value alignment, and meaningful human control over artificial intelligence. In A. Placani & S. Broadhead (Eds.), *Risk and Responsibility in Context* (1st ed., pp. 191–213). Routledge. <https://doi.org/10.4324/9781003276029-14>
- Oimann, A. K. (2023). The Responsibility gap and LAWS: A Critical mapping of the debate. *Philosophy & Technology*, 36(1), 3. <https://doi.org/10.1007/s13347-022-00602-7>
- Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.
- Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A., & Przybocki, M. A. (2021). *Four principles of explainable artificial intelligence (NISTIR 8312)*. National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.IR.8312>
- Regulation (EU) 2024/1689. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
- Retzlaff, C. O., Angerschmid, A., Saranti, A., Schneeberger, D., Röttger, R., Müller, H., & Holzinger, A. (2024). Post-hoc vs ante-hoc explanations: XAI design guidelines for data scientists. *Cognitive Systems Research*, 86, 101243. <https://doi.org/10.1016/j.cogsys.2024.101243>
- Ridzuan, N. N., Masri, M., Anshari, M., Fitriyani, N. L., & Syafrudin, M. (2024). AI in the financial sector: The line between innovation, regulation and ethical responsibility. *Information*, 15(8), 432. <https://doi.org/10.3390/info15080432>
- Robbins, S. (2019). A misdirected principle with a catch: Explicability for AI. *Minds and Machines*, 29(4), 495–514. <https://doi.org/10.1007/s11023-019-09509-3>
- Rudin, C., & Radin, J. (2019). Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review*, 1(2), 1–9. <https://doi.org/10.1162/99608f92.5a8a3a3d>
- Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems*, 7(1), 2400304. <https://doi.org/10.1002/aisy.202400304>
- Selbst, A. D., & Barocas, S. (2018). The Intuitive Appeal of Explainable Machines. *Fordham Law Review*, 87(3), 1085–1139.

- Slack, D., Hilgard, S., Singh, S., & Lakkaraju, H. (2021). Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in Neural Information Processing Systems*, 34, 9391–9404.
- Slota, S. C., Fleischmann, K. R., Greenberg, S., Verma, N., Cummings, B., Li, L., & Shenefiel, C. (2023). Many hands make many fingers to point: Challenges in creating accountable AI. *AI & Society*, 38(4), 1287–99. <https://doi.org/10.1007/s00146-021-01302-0>
- Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, 24(1), 62–77. <https://doi.org/10.1111/j.1468-5930.2007.00346.x>
- Srivastava, G., Jhaveri, R. H., Bhattacharya, S., Pandya, S., Rajeswari, Maddikunta, P. K. R., Yenduri, G., Hall, J. G., Alazab, M., & Gadekallu, T. R. (2022). *XAI for cybersecurity: State of the art, challenges, open issues and future directions*. arXiv. <https://doi.org/10.48550/arXiv.2206.03585>
- Stamatiou, F., & Karakonstantis, X. (2024). Should you trust your voice assistant? It's complicated, but no. *Linköping Electronic Conference Proceedings*. <https://doi.org/10.3384/ecp208015>
- Stepin, I., J. M. Alonso, A. Catala, & M. Pereira-Fariña. (2021). A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9, 11974–12001. <https://doi.org/10.1109/ACCESS.2021.3051315>
- Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy*, 48, 187–211.
- Tahir, H. A., Alayed, W., Hassan, W. U., & Haider, A. (2024). A novel hybrid XAI solution for autonomous vehicles: Real-Time interpretability through LIME–SHAP integration. *Sensors*, 24(21), 6776. <https://doi.org/10.3390/s24216776>
- Taylor, I. (2024). Is explainable AI responsible AI? *AI & Society*, 40(3), 1695–1704. <https://doi.org/10.1007/s00146-024-01939-7>
- Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Tamkin, A., Durmus, E., Hume, T., Mosconi, F., Freeman, C. D., ... Henighan, T. (2024). Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet [Research Report]. Anthropic. <https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html>
- Thompson, D. F. (1980). Moral responsibility of public officials: The problem of many hands. *American Political Science Review*, 74(4), 905–916. <https://doi.org/10.2307/1954312>
- Tigard, D. W. (2021). There is no techno-responsibility gap. *Philosophy & Technology*, 34(3), 589–607. <https://doi.org/10.1007/s13347-020-00414-7>
- Vale, D., El-Sharif, A., & Ali, M. (2022). Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law. *AI and Ethics*, 2(4), 815–826. <https://doi.org/10.1007/s43681-022-00142-y>
- Vallor, S., & Vierkant, T. (2024). Find the gap: AI, responsible agency and vulnerability. *Minds and Machines*, 34(3), 20. <https://doi.org/10.1007/s11023-024-09674-0>

- Van de Poel, I. (2015). The problem of many hands. In *Moral Responsibility and the Problem of Many Hands* (pp. 50–92). Routledge.
- Van der Waa, J., E. Nieuwburg, A. Cremers, & M. Neerincx. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291, 103404.
<https://doi.org/10.1016/j.artint.2020.103404>
- Vargas, M. (2013). *Building better beings: A theory of moral responsibility*. Oxford University Press.
- Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, 31, 841.
- Williams, I., Oldenburg, N., Dhar, R., Hatherley, J., Fierro, C., Rajcic, N., Schiller, S. R., Stamatou, F., & Søgaard, A. (2025). *Mechanistic interpretability needs philosophy*. arXiv.
<https://doi.org/10.48550/arXiv.2506.18852>
- Winner, L. (1978). *Autonomous technology: Technics-out-of-control as a theme in political thought*. MIT Press.
- Yadav, B. R. (2024). The ethics of understanding: Exploring moral implications of explainable AI. *International Journal of Science and Research (IJSR)*, 13(6), 1–7.
<https://doi.org/10.21275/SR24529122811>
- Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*, 34(2), 265–88. <https://doi.org/10.1007/s13347-019-00382-7>
- Zetsche, D. A., Arner, D. W., Buckley, R. P., & Tang, B. (2020). Artificial Intelligence in Finance: Putting the Human in the Loop. Centre for Finance, Technology and Entrepreneurship.
<https://ssrn.com/abstract=3531711>
- Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.