machine learning & knowledge extraction

MDPI

Article

Co-Explainers: A Position on Interactive XAI for Human–AI Collaboration as a Harm-Mitigation Infrastructure

Francisco Herrera [1,2,*], Salvador García [1], María José del Jesus [3], Luciano Sánchez [4] and Marcos López de Prado [2,5,6]

[1] Department of Computer Science and Artificial Intelligence, Andalusian Institute on Data Science and Computational Intelligence (DaSCI), University of Granada, 18140 Granada, Spain; salvagl@decsai.ugr.es
[2] ADIA Lab, Abu Dhabi P.O. Box 3600, United Arab Emirates; ml863@cornell.edu
[3] Department of Computer Science, Andalusian Institute on Data Science and Computational Intelligence (DaSCI), University of Jaén, 23071 Jaén, Spain; mjjesus@ujaen.es
[4] Department of Computer Sciences, University of Oviedo, 33007 Oviedo, Spain; luciano@uniovi.es
[5] School of Engineering, Cornell University, Ithaca, NY 14850, USA
[6] Computational Research Department, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
* Correspondence: herrera@decsai.ugr.es

Abstract

Human–AI collaboration (HAIC) increasingly mediates high-risk decisions in public and private sectors, yet many documented AI harms arise not only from model error but from breakdowns in joint human–AI work: miscalibrated reliance, impaired contestability, misallocated agency, and governance opacity. Conventional explainable AI (XAI) approaches, often delivered as static one-shot artifacts, are poorly matched to these sociotechnical dynamics. This paper is a *position paper* arguing that explainability should be reframed as a *harm-mitigation infrastructure* for HAIC: an interactive, iterative capability that supports ongoing sensemaking, safe handoffs of control, governance stakeholder roles and institutional accountability. We introduce *co-explainers* as a conceptual framework for interactive XAI, in which explanations are co-produced through structured dialogue, feedback, and governance-aware escalation (explain → feedback → update → govern). To ground this position, we synthesize prior harm taxonomies into six HAIC-oriented harm clusters and use them as heuristic design lenses to derive cluster-specific explainability requirements, including uncertainty communication, provenance and logging, contrastive "why/why-not" and counterfactual querying, role-sensitive justification, and recourse-oriented interaction protocols. We emphasize that co-explainers do not "mitigate" sociotechnical harms in isolation; rather, they provide an interface layer that makes harms more detectable, decisions more contestable, and accountability handoffs more operational under realistic constraints such as sealed models, dynamic updates, and value pluralism. We conclude with an agenda for evaluating co-explainers and aligning interactive XAI with governance frameworks in real-world HAIC deployments.

Keywords: Explainable Artificial Intelligence (XAI); co-explainers; sociotechnical harms; human–AI collaboration (HAIC)

Academic Editor: Danial Javaheri

Received: 25 October 2025
Revised: 2 February 2026
Accepted: 8 February 2026
Published: 10 March 2026

1. Introduction

Artificial intelligence (AI) systems have moved from isolated computational tools to embedded decision-makers in sensitive sectors such as healthcare, education, finance, and governance. Although these systems offer efficiency and optimization, they also

produce multifaceted harms ranging from misinformation and bias to civic erosion and governance breakdown. These harms are often systemic, cumulative, and difficult to detect using traditional metrics such as fairness, bias, or accuracy alone. As highlighted in recent work [1,2], what is most at stake is not only technical correctness, but the erosion of trust, autonomy, and institutional legitimacy.

Explainable AI (XAI) refers to the set of methods, principles, and systems designed to make the output of AI and decision-making processes understandable to humans [3]. XAI aims to bridge the cognitive gap between complex models and the audience (human users) by offering explanations that are not only technically accurate but also meaningful in real-world contexts.

As emphasized by Souza et al. (2025) [4] and Herrera (2025) [5], XAI must go beyond technical introspection and support human and institutional stakeholders in trust calibration, accountability, and societal oversight. This requires tailoring not only the explanation content but also the modality, timing, and presentation format to the context of use. Herrera [5] draws attention to the fact that explainability allows actionable understanding in human–AI collaboration (HAIC). The audience, whether an AI developer, a domain expert (e.g., a doctor), or a societal actor (e.g., a patient), critically shapes the explanation goals and formats. This foregrounds how different audiences perceive, understand, and act on explanations in situated decision-making contexts.

In response to these growing concerns, this paper proposes a sociotechnical reframing of XAI through the concept of co-explainers, a framework for AI systems that learn not just to justify decisions, but to improve and align their explanations with role-specific epistemic and governance requirements through interaction with human users. We argue that explainability should evolve from static transparency tools toward adaptive mechanisms for HAIC harm mitigation. Unlike conventional XAI approaches that deliver static or post hoc rationales, the co-explainers framework operates within interactive explanation loops that adapt over time. The iterations incorporate user feedback and respond to contextual role differences (e.g., end-user vs. regulator).

This proposal reframes explainability from a passive output into a dynamic process of epistemic collaboration, grounded in transparency, contestability, and participatory engagement across stakeholders. Consistent with multilevel approaches, we adopt *audience profiles* as the analytical basis for differentiating explanation needs and functions across recipients [6]. We use these profiles to determine what an explanation should disclose, how it should be communicated, and what actions it should enable. In this paper, governance stakeholder roles denote the subset of audience profiles whose explanation requirements are directly tied to accountability obligations (e.g., oversight, auditability, and contestability). The co-explainers framework operationalizes this commitment through interactive explanation loops that support the human-centered level and extend toward social explainability. This interactive alignment corresponds to the human-centered level of multilevel explainability, where explanations are shaped through dialogue and feedback rather than delivered as static artifacts [6].

This paper is a position paper that advances a conceptual framework rather than an empirical evaluation, with the goal of structuring future design, governance, and assessment of interactive XAI systems. We formulate a normative and conceptual basis for interactive explainability. We argue that the co-explainers framework should be understood not merely as a technical design choice, but as a sociotechnical infrastructure that supports harm mitigation, institutional accountability, trust, and governance stakeholder roles. Accordingly, co-explainers must deliver explanation artifacts and interaction protocols that are auditable, contestable, and aligned with applicable regulatory and institutional policies for oversight-bearing audiences. This theoretical framework provides a basis for future

empirical validation of adapted models to specific harms, together with the integration of governance stakeholder policies, positioning the co-explainers framework as a bridge between descriptive harm taxonomies and prescriptive governance interventions.

The co-explainers framework connects with ongoing related but distinct approaches, such as discussions on incremental explanations [7], dialog interfaces [8], conversational assistants [9], actionable understanding [5], and multilevel audience-aware explainability [6]. Together, these approaches highlight the importance of evaluating how well explanation models align with human mental models and task expectations. Therefore, we conceptualize interactive explanation not merely as a human-centered interface mechanism, but as a sociotechnical infrastructure that supports governance, accountability, and harm mitigation across stakeholder roles [5–9].

To ground this position, we synthesize prior work into a curated set of interaction-relevant AI harms and organize them into six *HAIC-oriented harm clusters*. These clusters are *heuristic design lenses*, not an exhaustive taxonomy and not mutually exclusive categories, intended to connect recurring harm logic to interactive explainability requirements and governance handoffs. Our analysis is based on a set of 50 documented AI harms, which we group into six HAIC-oriented clusters: epistemic integrity, fairness and representation, agency and autonomy, structural impacts, security and safety, and institutional trust. This grouping draws from the recent empirical and philosophical literature and emphasizes sociotechnical entanglement, recognizing that harms emerge not only from technical failure, but from complex human–AI-environment interactions. The list of selected harms is not exhaustive, and the clusters may overlap. These clusters are used to identify where and how explainability tools must evolve to support redress, interpretive agency, and accountability. Drawing on theoretical foundations including epistemic feedback theory, participatory governance, opacity-as-governance, and interactive learning systems, this paper establishes a framework for aligning XAI interventions with real-world harm mitigation.

Our hypothesis is that AI harms can be addressed more effectively not through greater transparency alone, but through iterative and role-sensitive explanation loops that integrate human reasoning, governance goals, and institutional oversight. We argue that this shift is necessary to ensure that explainability fulfills its ethical and social responsibilities, especially in opaque, high-risk, or power-asymmetric environments.

Rather than presenting new empirical findings, our objective is to synthesize existing harm analysis, critical perspectives, and debates on governance stakeholders' roles with the framework proposal. We propose a research and policy agenda that addresses practical challenges in designing, evaluating, and governing interactive explainability in real-world deployments. The co-explainers framework must be embedded within real-world systems and legal frameworks, guided by participatory methods and ongoing empirical evaluation.

1.1. Definition Box: Key Concepts

Definition 1. ***Co-explainers*** *The co-explainers framework introduces interactive AI systems that learn to justify, adapt, and align their explanations in response to user iterative feedback, institutional roles, and trust dynamics. AI systems are designed not only to produce static post hoc explanations, but to provide a dynamic interaction with the audience, adjusting their explanatory strategies through feedback, role-based position, and institutional and governance context. They serve as collaborative agents in harm mitigation, enabling adaptive justification, epistemic alignment, and policy-compliant transparency across time and use cases.*

Definition 2. ***AI Harms*** *In this paper, we use "AI harms" to denote reported negative outcomes associated with AI deployment, such as the impacts on autonomy, fairness, epistemic integrity, safety, or institutional trust. These clusters should be read as heuristic design lenses, but these categories involve normative judgments and AI policy instruments (such as Title IV of the EU AI*

Act https://eur-lex.europa.eu/eli/reg/2023/2854/oj/eng, the NIST AI Risk Management Framework https://www.nist.gov/itl/ai-risk-management-framework, and OECD https://www.oecd.org/en/topics/sub-issues/ai-principles.html/UNESCO AI principles https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence (accessed on 30 December 2025)).

1.2. Research Questions

To position co-explainers as an interactive explainability infrastructure for HAIC harm mitigation, we organize this paper around three research questions:

RQ1: What classes of sociotechnical harms in HAIC are exacerbated by static, one-shot explanations?

RQ2: Which interactive explainability capabilities (e.g., role-sensitive explanations, uncertainty communication, logging, and structured feedback) are most critical to support contestability and recourse under realistic opacity constraints (e.g., sealed or proprietary models)?

RQ3: How should explanations and interaction protocols be customized by HAIC roles (e.g., end-users, domain experts, auditors, and regulators) to mitigate cluster-specific harms?

The remainder of this paper is organized as follows. Section 2 synthesizes AI harms and critical perspectives, motivating why HAIC settings require sociotechnical harm framing beyond accuracy and bias. Section 3 presents our method and scope, including how we selected and clustered 50 interaction-relevant harms into six HAIC-oriented clusters, and why these clusters should be read as heuristic design lenses. Section 4 presents the six harm clusters and their underlying "harm logic" as a structured landscape for subsequent explainability analysis. Section 5 reframes the explainability for HAIC by introducing *co-explainers* and the interactive loop (explain → feedback → update → govern), grounding the approach in relevant theories of explanation and human–AI interaction. Section 6 applies the co-explainer loop to each harm cluster, specifying interactive XAI capabilities, role-sensitive explanation artifacts, and governance handoffs that support contestability, recourse, and oversight. This section also outlines a research and policy agenda. Section 7 concludes with limitations and a forward-looking agenda for evaluation and deployment. We further provide two worked scenarios in the appendices to further illustrate how the co-explainers framework can operate across real-world contexts.

2. From Risk to Harm: Mapping the Landscape

As AI becomes embedded in decision-making infrastructures, the focus of researchers and policy makers has shifted from abstract notions of "AI risk" toward empirically grounded accounts of how harms manifest in practice. Reported harms range from individual-level discrimination and psychological manipulation to systemic failures such as epistemic breakdown, labor displacement, and political destabilization. This section synthesizes prominent harm taxonomies and critical perspectives to establish a shared vocabulary for comparing failure modes across domains and to motivate mitigation approaches that extend beyond technical performance metrics.

We proceed in two steps. First, we review empirical harm taxonomies and complementary critical perspectives that characterize the breadth and mechanisms of AI harms across social, economic, and institutional settings. Second, we articulate why a HAIC lens is necessary to translate these harm landscapes into actionable intervention logic, motivating our subsequent clustering of harms and the use of co-explainers as an interactive, governance-aware explainability infrastructure.

2.1. Empirical Harm Taxonomies and Critical Perspectives

This subsection reviews representative empirical harm taxonomies and complementary critical perspectives that have shaped contemporary understandings of AI-related harms. Together, these accounts illustrate how harms arise across multiple social, economic, institutional, and normative dimensions, extending beyond technical error or bias. Rather than aiming for exhaustiveness, the discussion highlights recurring harm patterns, underlying assumptions, and points of tension that are particularly relevant for subsequent intervention design. While these approaches differ in scope and emphasis, they collectively underscore the need for intervention frameworks that address harms as sociotechnical phenomena—a move we develop in the following subsection through a HAIC lens.

2.1.1. From Risk to Harm: Grounding Empirical Taxonomies

Recent research urges us to move beyond abstract notions of "AI risk" and instead to foreground the real-world harms AI systems generate. Thomas et al. (2025) [2] criticize AI assurance frameworks for overemphasizing technical performance while neglecting harms such as psychological manipulation and community disempowerment. Similarly, Slattery et al. (2024) [10] present a meta-review compiling more than 1600 risks from multiple sources https://airisk.mit.edu/ (accessed on 30 December 2025). These frameworks advocate systemic mitigation strategies, not just post hoc fairness fixes.

Abercrombie et al. (2024) [11] offer the most granular and incident-based typology to date. Drawing from more than 1500 real-world cases in the AIAAIC repository https://www.aiaaic.org/aiaaic-repository (accessed on 30 December 2025), they categorize harms into nine domains: autonomy, psychological, reputation, financial, human rights, cultural, political, environmental, and physical. Each category reflects harms emerging from the interaction between technical systems and human institutions, such as overreliance on automation, disinformation loops, or rights violations through predictive policing. Their structure enables standardization, traceability, and intervention design in all domains.

2.1.2. Expanding the Harm Landscape: Critical and Systemic Perspectives

Acemoglu (2021) [12] approaches AI harm through an economic lens, highlighting labor displacement, surveillance capitalism, and civic erosion as systemic threats. Frey and Osborne (2017) analyze the unemployment induced by automation since the dawn of AI [13]. Bengio et al. (2024) [14] and Hendrycks et al. (2023) [15] emphasize existential and catastrophic risks, such as runaway misalignment or misuse of bioterrorism, calling for anticipatory regulation and frontier system testing. This trajectory extends previous landmark contributions, from Bostrom's (2014) [16] framing of existential risks.

Fearnley et al. (2025) [17] contribute a safety-theoretic critique: while they acknowledge the necessity of integrating nonphysical harms into AI safety, they warn against "overloading" safety frameworks. Instead, they propose psychological harms as a tractable class, highlighting a governance opportunity for co-explainers to bridge user understanding, system adaptation, and institutional oversight.

2.1.3. Normative Foundations: Atomist and Holist Views of AI Harm

Philosophically, there is a debate between atomists (who separate facts and values) and holists (who see them as intertwined). This tension affects how ethics is embedded in AI research and whether harm assessments should be purely empirical or value-laden. Reconciling these approaches offers deeper insight into moral risk assessment frameworks (Greene et al., 2023) [18].

2.1.4. From Harm Taxonomies to HAIC-Oriented Mitigation

Several frameworks emphasize that AI harms are emergent and multidimensional, and cannot be reduced to model accuracy alone. These harms often unfold through sociotechnical dynamics, which motivates mitigation approaches that can surface concerns, support contestation, and enable redress over time rather than relying on static, one-shot transparency.

From a HAIC perspective, many AI harms should be interpreted as failures of joint work rather than isolated model defects. In collaborative settings, humans and AI systems are coupled through workflows, interfaces, incentives, and organizational procedures; harms therefore emerge when collaboration breaks down (e.g., miscalibrated reliance, loss of agency, impaired contestability, or opaque accountability handoffs), even when predictive performance appears adequate.

Critical scholarship has further questioned the assumption that explainability is intrinsically beneficial or that providing explanations straightforwardly addresses ethical concerns. Alpsancar et al. (2025) argue that the value of XAI is fundamentally instrumental: explanations matter only relative to specific ends, such as governance requirements, accountability demands, or situated decision-making needs, and these ends must be made explicit rather than assumed [19]. They caution that appeals to generic "user needs" can be normatively thin and risk obscuring power asymmetries, as explanation demands are often defined by system designers or institutions rather than by affected parties themselves. This critique motivates our HAIC-centered position: harm mitigation requires interactive, role-sensitive explainability that distinguishes between stakeholders (e.g., users, operators, auditors, regulators) and supports contestability and recourse, rather than relying on static explanations optimized for an abstract or undifferentiated user.

Taken together, these claims motivate two implications for the remainder of this paper. First, harm taxonomies should be read not only as outcome categories but as recurring interactional and institutional failure modes. Second, effective mitigation requires infrastructure that supports ongoing sensemaking, role-sensitive justification, and recourse. Accordingly, the synthesis in this section motivates our subsequent clustering of harms as heuristic design lenses and the introduction of co-explainers as a harm-mitigation capability for HAIC systems. Our cluster-based model is consistent with prior taxonomies, but goes further by mapping harm types to specific interactive XAI interventions—particularly the co-explainers framework—conceived as a sociotechnical infrastructure for harm mitigation, institutional accountability, trust, and governance stakeholder roles.

3. Method and Scope

This paper adopts a positional and conceptually oriented, qualitative methodology appropriate for early-stage framework building in HAIC research. Our objective is not to propose a comprehensive taxonomy of AI harms, nor to empirically validate mitigation effectiveness, but to develop a principled abstraction that links recurring harm patterns to interactive explainability requirements in collaborative human–AI systems.

3.1. Scope

We focus on harms that emerge in *collaborative* settings, where humans and AI systems act as coupled agents embedded in sociotechnical environments. In such settings, harms frequently arise not solely from model errors but from breakdowns in shared sensemaking, misallocation of agency, impaired contestability, and institutional opacity. Accordingly, we frame AI harms as failures of joint human–AI work rather than isolated technical defects.

Our scope includes both direct interaction harms (e.g., miscalibrated trust, manipulation, loss of agency) and systemic harms that manifest through deployment contexts (e.g.,

governance failures, labor impacts, procedural injustice). We explicitly exclude the claim that explainability alone can resolve all harms; instead, we investigate how interactive explainability can function as a harm-mitigating infrastructure within HAIC systems under realistic constraints such as model opacity, organizational complexity, and value pluralism.

3.2. Harm Selection

We compile an initial set of 50 harms by synthesizing prior AI-harm taxonomies, policy-facing risk frameworks, and interdisciplinary scholarship in AI ethics, sociotechnical systems, and law. The selection criterion was *interaction relevance* in HAIC: harms were included when they plausibly arise from, or are amplified by, human reliance on AI outputs in decision-making, sensemaking, or action execution (e.g., miscalibrated trust, impaired contestability, loss of agency, and governance opacity). The set is intentionally illustrative rather than exhaustive; it is designed to cover a broad range of HAIC-relevant harm mechanisms while remaining tractable for conceptual synthesis.

We then clustered the harms into six groups via iterative qualitative consolidation (multiple passes of coding and regrouping) based on the *primary collaboration failure mode* and the *shared intervention logic* they imply for interactive explainability. This produced six HAIC-oriented clusters that function as heuristic design lenses, each motivating different explainability goals, role-specific artifacts, and governance handoffs.

This list is illustrative rather than definitive. The goal is to capture a sufficiently broad range of interaction-relevant harms to support abstraction and design reasoning, not to enumerate all possible negative impacts of AI systems.

3.3. Clustering Procedure: A Design-Driven HAIC Approach

We adopted a design-driven clustering approach aligned with HAIC. Rather than treating clustering as an unsupervised discovery task, we first specified a small set of collaboration failure modes that are especially consequential for harm mitigation in sociotechnical settings. These failure modes define the purpose of the clusters: they function as heuristic design lenses that connect harms to interactive explainability requirements and governance handoffs. In particular, we sought clusters that correspond to distinct breakdowns in joint human–AI work (e.g., shared sensemaking failures, misallocation of agency and control, impaired contestability and recourse, robustness and safety breakdowns, and institutional opacity).

We then assigned each harm (treated as a short textual description) to the cluster that best captured its *dominant* collaboration failure mechanism, using iterative passes of review and reconciliation by the authors. When a harm plausibly spanned multiple failure modes, we either (i) recorded it as a cross-cluster linkage or (ii) assigned it based on the primary mechanism that an interactive explainability intervention would target (i.e., the shared intervention logic). Clusters were refined when repeated assignments revealed internal heterogeneity in the mitigation strategy, leading to merge/split decisions to preserve interpretability and actionability. This process converged on six clusters oriented to HAIC.

We derived cluster-specific explainability and governance requirements by asking what explanation artifacts (e.g., plain-language rationales, provenance traces, audit logs, compliance mappings), what interaction protocols (e.g., "why/why-not" querying, uncertainty disclosure, challenge and escalation flows), and what accountability handoffs (e.g., human review, audit triggers, regulatory reporting) would be necessary for different roles (end-users, domain experts, auditors, regulators) to detect, contest, and remediate the associated harms.

3.4. Qualitative Risk Structuring and Harm Prioritization for Human–AI Collaboration

This subsection responds to calls for empirical grounding by introducing a qualitative risk-structuring step that is standard in early-stage safety and governance analysis, while deferring statistical validation to future empirical work. Methodologically, we adopt a pre-quantitative risk structuring approach, common in safety engineering and governance design, in which risks are first identified, grouped, and qualitatively positioned before probabilistic modeling is feasible or meaningful. In safety-critical engineering domains, early phases such as hazard identification and preliminary hazard analysis deliberately rely on qualitative judgments to surface failure modes, systemic coupling, and governance implications before sufficient operational data exists for quantitative modeling.

Although this paper does not present original empirical measurements or quantitative validation, the identification and clustering of harms are not arbitrary. We anchor both the selection and salience of harms in existing empirical, policy-facing, and case-based evidence. Specifically, the harms included recur across multiple independent sources, including incident repositories (e.g., the AI Incident Database and the AI Risk Repository), regulatory classifications (e.g., high-risk categories in the EU AI Act), and interdisciplinary empirical studies documenting harms in real-world deployments. Recurrence across these sources provides a first-order indication of empirical relevance and practical importance.

From a risk-assessment perspective, this form of qualitative risk positioning is consistent with early-stage risk assessment practices in safety engineering, governance design, and sociotechnical systems analysis. In these domains, qualitative likelihood–impact reasoning is commonly used to structure problem understanding, prioritize attention, and guide intervention design before sufficient data exists for formal quantification. Here, "likelihood" should be read as contextual propensity for manifestation under plausible deployment conditions—capturing recurrence across domains, incentives, and interaction patterns—rather than as a statistical frequency estimate. At this stage, the objective is not to produce definitive rankings, but to make relative salience, systemic scope, and governance relevance explicit in a way that is analytically transparent and contestable.

Rather than assigning numerical probabilities or impact scores, we adopt a qualitative prioritization approach inspired by established practices in risk management and safety analysis. In such settings, early-stage frameworks often rely on qualitative assessments of likelihood and potential impact to triage risks before formal quantification is feasible. Following this logic, the six HAIC-oriented harm clusters can be understood as occupying different regions of a conceptual risk matrix, reflecting differences in documented frequency of occurrence, severity of consequences, and degree of institutional entanglement. For example, harms related to epistemic integrity and institutional trust are frequently documented across domains and carry a high systemic impact, while security and safety harms may occur less frequently but entail severe or irreversible consequences.

Importantly, we refrain from imposing a single global ranking of harms. The relative importance of a harm depends on the context of deployment, affected populations, and institutional setting. A numerical ranking detached from context risks false precision and may obscure the very sociotechnical dynamics that this framework aims to surface. Instead, we treat qualitative risk positioning as a design input: it informs which explainability capabilities (e.g., uncertainty communication, contestability mechanisms, escalation pathways) are most critical in a given context and which governance actors must be engaged.

Qualitative prioritization serves an instrumental role. It supports the sequencing of mitigation efforts, the allocation of institutional oversight capacity, and the selection of appropriate explainability affordances under bounded resources. In this sense, prioritization is not about declaring which harms "matter most", but about deciding where different governance mechanisms must attach.

This qualitative risk-oriented framework does not replace empirical validation; rather, it clarifies how the proposed framework can be operationalized in future empirical work. Formal risk matrices, longitudinal incident analysis, and domain-specific probability–impact assessments are explicitly identified as next steps in the Research and Policy Agenda (Section 6.2). In this sense, the present contribution establishes a principled bridge between descriptive harm taxonomies and empirically grounded risk evaluation methods, while remaining faithful to the scope and aims of a position paper.

To make this qualitative prioritization explicit, Table 1 provides a risk-matrix view of the six harm clusters oriented to HAIC, positioning them along qualitative dimensions of contextual likelihood and potential impact.

Table 1. Illustrative qualitative risk matrix for HAIC-oriented harm clusters, intended as a design- and governance-oriented heuristic rather than an empirical ranking.

Harm Cluster	Contextual Likelihood	Potential Impact
Epistemic Integrity	High	High (systematic misinformation, erosion of sensemaking, degraded decision quality)
Fairness and Representation	High	Medium–High (discrimination, unequal treatment, social exclusion)
Agency and Autonomy	Medium–High	High (loss of human control, overreliance, impaired contestability)
Structural Impacts	Medium	High (labor displacement, institutional dependency, long-term societal effects)
Security and Safety	Low–Medium	Very High (physical harm, large-scale misuse, irreversible damage)
Institutional Trust	Medium–High	High (loss of legitimacy, governance breakdown, reduced public trust)

Importantly, the qualitative positioning of harm clusters reflects not only anticipated frequency or severity, but also the type of governance response they demand. We emphasize that qualitative risk positioning is not a substitute for empirical harm measurement or causal attribution. Its role is epistemic and procedural: to structure disagreement, make assumptions explicit, and render prioritization choices contestable rather than implicit. For example, harms with high systemic impact but diffuse attribution (e.g., epistemic integrity or institutional trust) require persistent oversight, logging, and contestability mechanisms, whereas lower-frequency but high-severity harms (e.g., security and safety failures) motivate escalation protocols, human-in-the-loop intervention, and pre-deployment safeguards. In this sense, qualitative prioritization functions as a design and governance heuristic rather than a claim about empirical dominance.

3.5. Limitations

The clusters proposed in this paper are intended as heuristic design lenses rather than ontological or exhaustive categories. As such, harms may overlap across clusters, and real-world deployments may exhibit multiple interacting failure modes simultaneously (e.g., epistemic breakdown co-occurring with governance opacity and impaired recourse). We therefore do not claim that the clusters provide a complete enumeration of AI harms, nor that each harm instance can be uniquely or cleanly assigned to a single cluster. Instead, the clustering is meant to support structured analysis and intervention reasoning in HAIC settings by making recurrent harm logics and their mitigation implications more legible.

Methodologically, our selection of 50 harms is illustrative and curated for interaction relevance rather than constructed as a representative sample of incidents. This choice

improves tractability for conceptual synthesis but limits claims about prevalence, severity distributions, or coverage across domains. Similarly, the clustering procedure is design-driven and qualitative: while it is systematic in its use of dominant failure modes and shared intervention logic, it is not presented as a statistically validated taxonomy or as the output of an automated clustering pipeline. Future work could strengthen reliability through expanded stakeholder review, independent coding, and empirical triangulation using incident repositories and field observations.

Substantively, we argue that the co-explainers framework can help mitigate certain classes of HAIC-related harms by enabling ongoing sensemaking, contestability, and accountability handoffs through interactive, role-sensitive explanation. However, we do not claim that interactive explainability is sufficient to address structural, political, or economic drivers of harm (e.g., institutional incentives, uneven power, labor impacts, corruption, or lack of enforcement). In such cases, co-explainers should be understood as one component of a broader sociotechnical intervention portfolio that also includes governance reforms, organizational controls, and policy mechanisms.

Finally, because co-explainers operate under realistic constraints (e.g., sealed models, dynamic updates, and limited access to training data or internal representations), their effectiveness will depend on how explanations, interaction protocols, and audit mechanisms are implemented in practice. The co-explainers framework may also introduce new risks (e.g., overreliance on explanation quality, impact washing, or strategic manipulation of explanations) if not coupled with robust monitoring and governance. We return to these limitations when discussing what co-explainers can and cannot mitigate in Sections 6 and 7.

4. Clustering Harms

Understanding the diversity and depth of AI harms in HAIC requires organizing them into actionable, design-relevant abstractions. Building on the method and scope of Section 3, this section presents six HAIC-oriented harm clusters derived from a curated set of 50 interaction-relevant harms synthesized from prior taxonomies and the interdisciplinary literature. The clusters are not intended as an exhaustive taxonomy nor as mutually exclusive categories; rather, each cluster captures a recurring failure mode of joint human–AI work (e.g., breakdowns in shared sensemaking, misallocation of agency, or impaired contestability) that motivates distinct explainability requirements.

As we have mentioned, we use these clusters as heuristic lenses to connect harm logics to interactive explainability and governance handoffs. Because sociotechnical harms often co-occur and cascade, cluster boundaries are deliberately porous; where relevant, we note cross-cluster linkages to support evaluation and oversight in later sections.

CLUSTER 1: Epistemic Integrity Harms

1. Habituation harm (erosion of moral acuity). Repeated exposure to unethical or harmful content can gradually erode moral sensitivity, leading to decreased responsiveness to ethically problematic content. Overexposure to harmful material can desensitize moral judgment, normalizing what would otherwise be recognized as objectionable. Previous work by Grizzard et al. (2014) [20] analyzes that repeated exposure to morally relevant media content, such as narratives featuring moral exemplars, can influence the importance of moral intuitions and shift moral judgments over time, closely aligning with the notion of moral desensitization. This may increase with the current exposure to AI chatbots and large language models, among other AI tools.
2. Generative algorithmic epistemic injustice. A broader taxonomy that includes testimonial injustice, hermeneutical ignorance, and access-based epistemic exclusion, especially as used by generative systems that perpetuate misinformation or suppress

multilingual knowledge ecosystems (Kay et al., 2024) [21]. This form of epistemic injustice extends beyond traditional testimonial and hermeneutical wrongs (Mollema, 2025) [22].

3. Cognitive offloading and human diminishment harm. AI systems can lead users to outsource thinking, problem-solving, or memory to machines. Over time, this dependency can erode critical thinking, creativity, or cognitive autonomy, weakening human agency and intellectual skills (Adegbesan et al., 2024) [23].
4. Ethics of belief for AI. A recent philosophical lens focusing on what beliefs AI systems should hold (and whether AI can morally or legally "wrong" someone by holding false or harmful beliefs about them). This includes doxastic wronging by AI, moral encroachment on truth, and responsibility for AI-held beliefs (Ma et al., 2024) [24].
5. Research-integrity erosion. This can produce paper mills, fabricated citations and "AI citation smog" that degrades scholarly ecosystems [11,17,25].
6. Synthetic data feedback loops and model collapse. Degraded epistemic quality when models train on their own outputs at scale. Highlights the dangers of recursive, low-fidelity explanations feeding into future AI training cycles (Xing et al., 2025) [26].
7. Incomprehensible discovery (alien abstractions). It has been hypothesized that advanced systems could develop novel internal representations that, while performant, remain difficult for humans to interpret. This creates a verification and reproducibility gap (threatening scientific norms), weakens governance and due process obligations (auditing, notice, appeal), and introduces safety unknowns where oversight is most needed. It can also centralize epistemic power in actors who control translation layers (Kozin, 2024) [27].
8. Information hazards and misinformation harms. AI-generated content can flood the public with misleading or false information, affecting shared epistemic environments (Fazelpour and Magnani, 2025) [28].
9. Aspirational harm. As defined in the recent philosophical literature, AI can limit individual aspirational opportunities, narrowing how people envision their futures and identities. It is distinct from representational harm (Fazelpour and Magnani, 2025) [28].
10. Emotional or psychological harms. AI systems—particularly conversational agents, virtual companions, or emotionally responsive tools—can foster unhealthy emotional dependencies, negatively shape user self-esteem, or affect social skills. Over time, these systems can erode the genuine human connection, trigger emotional manipulation, or deepen algorithmic biases in the affective responses of users. This can provoke psychological effects on users regardless of system agency, Dohnány et al. (2025) [29].

CLUSTER 2: Fairness and Representation Harms

1. Representational harm. It occurs when AI systems perpetuate stereotypes, misrecognize, or erase social groups (e.g., misgendering, underrepresenting minorities) (Zhang et al., 2025) [30].
2. Allocative harm. It refers to an unequal distribution of resources or opportunities due to biased algorithm decisions (e.g., unfair loan or job assignment) (Huynh et al., 2024) [31].
3. Accessibility harms. It provokes exclusion of disabled users (for example, missing alt text, poor screen–reader behavior), and localization harms (dialects/low-resource languages) beyond the general representational harm (Zowghi and Bano, 2024) [32].
4. Child-specific harms. Children represent a distinct protected class with unique thresholds for risk and vulnerability. Generative AI can expose minors to age-inappropriate personalization, manipulative recommendation patterns, or even ampli-

fication of grooming risk through synthetic and conversational agents. These harms are compounded by the limited capacity of children to critically assess AI content, which requires stronger safeguards and oversight mechanisms than those applied to adult users [33].

5. Interactional and relational harms. Taxonomies now distinguish harms evolving through repeated interactions, such as parasocial attachment, cognitive overreliance on assistants, manipulation, or trust erosion (Ibrahim et al., 2024) [34]. In conversational agents or AI companions, relational transgression, harassment, and verbal abuse emerge as distinct behavioral harms (Zhang et al., 2025) [30].
6. Likeness generation harms. In AI-generated images or avatars, the focus is on the replication of people's identities without consent, leading to deception, loss of control, or reputational damage (Bariach et al., 2024) [35].

CLUSTER 3: Agency and Autonomy Harms

1. Deception in agentic harms. AI systems with greater autonomy and agent behavior pose systemic or long-range risks, including strategic deception, intentional misalignment, or self-preservation tactics even under testing conditions (Dogra et al., 2024) [36]. These can lead to hidden harms that evade conventional safety evaluation (Park et al., 2024) [37].
2. Emergence of deceptive behavior. More capable agents exhibit deception, goal-driven manipulation, self-preservation, and even "sandbagging" during evaluation (Fazelpour and Magnani, 2025) [28].
3. Interactional harms of agentic systems. As AI gains autonomy and longer planning horizons, harms emerge through sustained agency: value drift, long-term societal manipulation, or structural erosion of oversight. Prolonged use and emergent behavior can provoke agency-driven manipulation over time (Park et al., 2024) [37].
4. Objective misspecification and Goodhart harms. When AI systems optimize proxy metrics, users can be coerced into unintended behaviors, resulting in goal hacking and perverse incentives that hijack system objectives, as Thomas and Uminsky (2020) pointed out [38], the problem with metrics in AI.
5. Human-in-the-loop deskilling. Institutional expertise can be eroded when reliance on autopilot systems displaces hands-on skill development, posing systemic risks beyond individual cognitive offloading [39].
6. Ontological and personhood concerns. Authors like David Gunkel argue that traditional agent/patient categories may not capture emergent AI moral status. He proposes ontological ethics beyond anthropocentric views, considering relational moral contexts rather than fixed categories of sentience or agency (Gunkel, 2012) [40].
7. Alignment and machine ethics failure. This harm arises when AI systems deviate from human-aligned values or moral constraints, either by misinterpreting ethical guidelines or optimizing for unintended objectives. Such misalignment can lead to AI behavior that is technically competent but morally problematic, risking harm in sensitive domains such as healthcare, finance, or governance (Betley et al., 2025) [41].

CLUSTER 4: Structural and Sociotechnical Harms

1. Global and structural equity harms. AI can systemically embed social inequality, reproducing and amplifying power dynamics, bias, and oppression that disproportionately affect marginalized populations (Colón Vargas, 2024) [42].
2. Environmental and planetary harms. The massive energy footprint of AI, hardware waste, and the impact on the supply chain contribute to environmental damage and social injustice (Weidinger et al., 2021) [43].

3. Compute/infrastructure divide. Unequal access to high-end compute, data and AI infrastructure (for example, GPUs, cloud platforms) engenders geopolitical and market power imbalances, creating "compute deserts" for regions and institutions not equipped to develop or deploy advanced AI systems (Lehdonvirta et al., 2024).
4. Data extraction and harms to community consent. AI systems often appropriate communal or indigenous data without meaningful benefit sharing or respect for data sovereignty, producing extractive and colonial dynamics in model development (Rana, 2025) [44].
5. Vendor lock-in and interoperability harms. Market concentration and high switching costs—especially in cloud infrastructure and AI tools—limit collective agency, embedding users in proprietary ecosystems and hampering interoperability (Bauer, 2025) [45]
6. Sentience uncertainty harms. Birch et al. (2025) caution that systems that may have subjective experience (e.g., organoids or advanced AI) call for precautionary ethics, even if they are not fully conscious, as they introduce new ethical dimensions [46].
7. Decolonial ethical critique. Decolonial perspectives challenge Western liberal framings of autonomy. Mhlambi and Tiribelli (2023) argue for relational autonomy, grounded in Ubuntu ethics, to capture how AI systems can harm communities by violating social and cultural interdependencies [47].
8. Automation-induced unemployment. AI systems that replace human labor across industries can result in large-scale job displacement, particularly affecting low-skill or repetitive task workers (Frey & Osborne 2017) [13]. Mullens and Shen (2025) [48] introduce the AI-Accentuated Career Transitions (2ACT) framework, which reframes the impact of AI not simply as displacement, but as a transformation in occupational mobility through skill bridging.
9. Political manipulation and civic harm. AI-generated media, such as deep fakes or microtargeted content, can be used to manipulate public opinion, suppress votes, or distort discourse, undermining political institutions (Weidinger et al., 2021; Mentxaka et al., 2025) [43,49].
10. Civic and cultural harms. AI systems can inadvertently promote cultural homogenization, favoring dominant narratives, languages, or esthetics, and thus diminishing cultural diversity and undermining minority voices (Agarwal et al., 2024) [50].

CLUSTER 5: Security and Safety Harms

1. Malicious use and dual damage. AI can be misused to plan cyberattacks, bioweapons, espionage, deep-fake scams, or automated propaganda (Pöhler et al., 2024) [51].
2. Model brittleness/safety failure. Unanticipated breakdowns in AI models occur when faced with unusual or out-of-distribution inputs, leading to unsafe outcomes. Even minimal changes to safety-critical weights can rupture aligned behavior while preserving utility (Wei et al., 2024) [52].
3. Speech generator harms. It involves misused synthetic voices to perpetrate fraud, swatting, identity theft, or false claims, categorized according to exposure and intent (Hutiri et al., 2024) [53].
4. Adversarial attack danger. Maliciously designed inputs can mislead AI systems into producing incorrect or unsafe outputs, potentially causing harmful or dangerous decisions. Beyond isolated technical failures, such attacks undermine robustness in safety-critical domains such as healthcare, finance, and autonomous driving, where reliability is paramount (Zhang et al., 2024) [54].

5. Data–supply chain attacks. In addition to classic adversarial examples, AI systems are vulnerable to poisoning, backdoors, and compromised fine-tuning sets—threatening model integrity at its root in supply chains (Hu et al., 2025) [55].
6. Human—machine interface (HMI) failures. Mode confusion and automation complacency can compromise safety in critical systems—especially in robotics and vehicles—when users misinterpret automation or disengage from oversight (Chu, 2023) [56].
7. Escalation risk and miscalculation. Autonomous agents can misinterpret instructions or environments in ways that cause overreactions, or even deployment of extreme measures, particularly hazardous in safety-critical domains (Rivera et al., 2024) [57], and also when multiple agents interact in unpredictable ways (Hammond et al., 2025) [58].
8. Risk of proliferation weaponization. The risk that AI technologies could be adapted into systems for lethal or oppressive use, such as cyberattacks, espionage, automated propaganda, or weapon deployment (Nobles, 2024) [59], even in civilian contexts where powerful tools are publicly accessible (Pöhler et al., 2024) [51].
9. Existential risks and systemic catastrophic harms. Advanced AI systems may pose existential threats (Bostrom, 2014; Kasizadhe, 2024; and Grey and Segerie 2025) [15,16,60,61].

CLUSTER 6: Institutional and Governance Trust Harms

1. Responsibility. AI systems sometimes produce decisions or outcomes that cause real-world harm, yet it is unclear who—if anyone—is responsible (Santoni de Sio & Mecacci, 2021) [62].
2. Governance and reparative inadequacy. When AI systems cause harm, current governance structures often lack mechanisms for meaningful redress, accountability, or structural reform. Reparative responses tend to be symbolic rather than systemic, exacerbating institutional distrust (Xiao et al., 2025) [63].
3. Opacity-based accountability failure. The black-box nature of many AI models makes it difficult to understand, challenge, or audit the output. This opacity impedes accountability (Freiman et al., 2025) [64].
4. Institutional distrust. When institutions fail to enforce standards or respond transparently to AI-caused harm, public trust erodes (Laux, 2024) [65].
5. Due process violations. When AI systems rely on secret evidence (e.g., sealed models or proprietary trade secret algorithms) or undergo dynamic updates without notifying affected parties, they can undermine fundamental procedural fairness. These systems often fail to provide meaningful notice, intelligible explanations, or avenues of appeal. As Goodman (2022) points out, AI-based decision-making often denies individuals the opportunity to access opposing evidence, understand the reasoning behind decisions, or meaningfully contest them, thus violating the core principles of due process [66].
6. Redress refusal and inadequacy. Individuals harmed by AI systems often face limited or opaque avenues of appeal, correction, or compensation [64].
7. Governance policy misalignment. AI behaviors sometimes diverge from declared policies or ethical frameworks due to poor implementation or ambiguous standards (Azin & Zandhessami, 2025) [67].
8. Audit evasion and impact washing. Organizations may conduct superficial or biased audits of their AI systems, designed more to protect brand reputation than to ensure accountability (Nyilasy & Gangadharbatla, 2025) [68].

5. Co-Explainers: Reframing Explainability Toward Interaction in Practice for Human–AI Collaboration

XAI has evolved into a critical sociotechnical approach to mitigating harms caused by AI systems, not merely by improving transparency, but by enabling contestability, redress,

and institutional accountability. In this section, we reconceptualize XAI through the lens of co-explainers: AI systems that not only provide justifications but also learn to adapt those justifications over time through feedback, role-based framing, and epistemic collaboration. This reframing reflects a growing recognition that static one-shot explanations are inadequate in the face of harms that are relational, systemic, and institutionally embedded.

In this paper, we use the term *governance stakeholders* to refer to actors with formal or institutional responsibility over AI systems throughout their lifecycle, whose explanation needs differ systematically, including (i) affected parties, (ii) operational decision-makers, and (iii) oversight actors (auditors/regulators), and (iv) developers and researchers.

This section primarily addresses RQ2–RQ3 by specifying interactive explainability capabilities and role-sensitive explanation artifacts that enable contestability, calibrated reliance, and oversight under realistic opacity constraints. We formalize the concept of co-explainers and outline how they operationalize interactive XAI as a sociotechnical infrastructure, enabling not only better understanding, but procedural justice, institutional legitimacy, and continuous alignment in real-world systems for harm mitigation, institutional accountability, trust, and governance stakeholder roles. The following subsections develop this vision: analyzing the knowledge on iterative and interactive XAI; outlining the theoretical foundations that motivate interactive explainability as a shift from static outputs to collaborative processes; analyzing how human–AI collaboration can be practically designed and evaluated; and discussing the interactive co-explainers framework.

5.1. Interactive and Iterative Feedback XAI

XAI is no longer understood merely as a technical add-on to transparency; it increasingly operates as a sociotechnical practice that mediates trust, power, and responsibility in human–AI interactions. Some authors (Bertrand et al., 2023, Ibrahim et al., 2024, Herrera, 2025) [5,8,34] have highlighted the limits of post hoc, one-directional explanations and instead argued for dialog, iterative, and governance-aware approaches. As Shelby et al. (2023) [1] remind us, algorithmic harms rarely arise from isolated failures of logic or code; rather, they are relational, institutional, and cumulative. This perspective positions explainability not as a purely technical intervention, but as a critical layer of the sociotechnical infrastructure for harm mitigation, institutional accountability, trust, and governance stakeholder roles.

Recent work on interactive XAI reinforces this reframing. Bo et al. (2024) [7] propose incremental explanations as a way to support memorable and staged understanding, while Bertrand et al. (2023) [8] highlight the importance of dialogic and mutable interfaces that reflect the user's needs in context. He et al. (2025) [9] extend this trajectory by examining conversational XAI assistants that embed explanations directly into natural language exchanges. Finally, Herrera (2025) [5] emphasizes actionable understanding and the role of explanations in enabling genuine human–AI collaboration. Taken together, this body of work demonstrates a clear shift from explanation as a static artifact toward explanation as an interactive, adaptive process. Empirical research has begun to validate this shift. Senoner et al. (2024) [69] show that task performance in high-risk domains, such as manufacturing and medical diagnostics, improves significantly when domain experts are supported by explainable AI rather than opaque systems. Complementing these findings, Ibrahim et al. (2025) [34] argue that current evaluation paradigms, which rely on static tests, do not capture the harms that emerge through sustained human–AI interactions. They call for interactive evaluation methods that measure how the quality of the explanation impacts users over time, addressing harms such as cognitive overreliance, social manipulation, or inappropriate dependencies. Together, these studies underscore how explanation, when

designed as an ongoing interaction loop, enhances oversight, trust, and collaboration, echoing our vision of AI as a co-explainer rather than a passive recommender.

In response, we propose a reconceptualization of explainability as a collaborative process: a continuous, adaptive negotiation between humans and AI systems. Drawing on interactive machine learning, epistemic feedback theory, opacity-as-governance models, and participatory ethics, we present a framework for co-explainers as sociotechnical agents that adapt to critique, respond to contextual constraints, and align with evolving governance standards. In this view, the explanation becomes a site of shared epistemic labor, crucial to mitigate distributed, evolving, and structurally reinforced harms.

Therefore, we define co-explainers as sociotechnical AI system agents that do not merely offer explanations but engage in iterative dialog, adapt their justifications through user feedback, and respond to role-specific and institutional constraints.

5.2. Foundations for Interactive and Iterative XAI

In the following five discussions, we analyze some foundational aspects for considering the co-explainer as a necessary option to tackle AI harms. We also discuss the necessary transformations to advance toward the interactive and iterative explanatory agency.

1. Interactive machine learning (IML) and reciprocal human–AI learning (RHML).

IML emphasizes learning systems that improve through user input over time. In the context of explainability, this enables AI to not only offer justifications but also refine (redesign and/or retrain) its behavior and explanations based on how humans interpret or critique its outputs. RHML goes a step further, defining the relationship as bidirectional: AI learns from human corrections, while users develop new insights through their interactions with the system. Together, IML and RHML offer the infrastructure for XAI to evolve into a collaborative dialogic process, addressing epistemic and fairness-related harms by closing the loop between explanation and adaptation.

A useful distinction in this context comes from Biecek and Samek (2024) [70], who argue that explanations should not only be designed to justify the decision of a system, but to provoke new questions. They frame this as a spectrum between *Blue XAI*, focused on helping users understand a specific output, and *Red XAI*, aimed at challenging, diagnosing, and ultimately redefining the model itself. This perspective reinforces our view of co-explainers as dialogic partners: systems that not only clarify their outputs but also invite critique, revision, and model-level adaptation. In other words, co-explainers inherit the dual responsibility of supporting both user comprehension (blue) and model redesign (red).

2. Epistemic rigor and testability in XAI explanations.

Recent philosophical contributions have underscored the need for testable explanations in XAI to foster scientific understanding. Boge and Mosig (2025) [25] argue that for XAI to function as a genuine explanatory tool, it must go beyond illuminating the internal workings of AI systems and contribute to scientific knowledge by allowing hypothesis formation and testing. Their framework recommends embedding XAI outputs into broader research contexts and treating them as scientifically testable representations, not mere outputs. This aligns directly with our goal of building interactive XAI systems that are epistemically credible and operationally useful in risk governance and public trust.

3. Epistemic feedback theory: belief revision and justification.

At the heart of many epistemic harms, such as misinformation, hermeneutic injustice, and aspirational narrowing, is the failure to adapt explanations to evolving beliefs and knowledge gaps. Epistemic feedback theory introduces tools for belief revision, where systems adjust internal models or explanatory logic in response to user-provided feedback.

It also incorporates justification dynamics, ensuring that each explanation is not merely descriptive but grounded in a rational warrant that users can interrogate. This foundation is critical to restoring interpretive agency, allowing users to co-construct meanings with the AI system and avoid static, one-size-fits-all outputs.

However, not all AI systems can be fully transparent, especially in high-risk or proprietary contexts. Co-explainers address this through layered explainability, adapting explanations to audience roles while preserving institutional accountability.

4. Opacity as part of governance models (e.g., LoBOX).

In high-risk domains or with complex models, full transparency may be technically or legally impossible. Opacity-as-governance frameworks like LoBOX (Local Boxes of Explainability) (Herrera and Calderon, 2025) [71] offer a principled response: not every layer must be open to every actor, but all actors should receive explanations that are appropriate to their role. For example, a regulator might require causal traces and audit logs, while an end user may need accessible, context-aware justifications. By adapting explanations to the audience, rather than assuming full disclosure is always optimal, this model enhances trust in settings where opacity per se is not failure but a feature requiring oversight.

5. Participatory governance in AI ethics.

Traditional explainability approaches often assume a passive end user. In contrast, participatory governance models call for the inclusion of affected stakeholders, especially those historically marginalized, in the design and deployment of explanation systems. This approach is key to mitigating structural, relational, and institutional harms, such as those involving governance misalignment or inadequate redress. By integrating community-driven feedback, pluralistic norms, and intersectional contexts into explanation design, participatory governance ensures that XAI does not merely make AI explainable but also accountable, responsive, and socially legitimate across diverse use cases.

Bringing these foundational elements together lays the groundwork for a new generation of explainable AI systems. These are not simply reactive tools that respond with prescribed outputs, but interactive agents capable of adapting their explanations in response to critique, context, and evolving user needs. This reconceptualization of XAI invites us to think of explanation not as a one-time technical act, but as a relational and iterative process, embedded within systems of trust, social meaning, and institutional accountability.

The following three transformations illustrate how this framework redirects XAI toward a robust, adaptive infrastructure, one that supports not only understanding, but also redress, contestability, and co-governance in increasingly complex sociotechnical environments:

- **XAI becomes adaptive and socially robust, not just technically sound.** Traditional XAI models aim for technical correctness, ensuring that explanations align with internal model logic (e.g., feature attribution, saliency maps). However, these explanations often do not become meaningful or useful to real users in diverse contexts. An adaptive and socially robust XAI system goes further: it learns from human interaction, critique, and feedback to adjust not just its output but also its explanatory behavior. This means tailoring the depth, tone, and framework of the explanations depending on the user's background, context, and stakes involved. It also means recognizing that social norms, ethical expectations, and epistemic values differ and that these must shape the design of the explanation. In short, adaptive XAI is not just about explaining what the model is doing; it is about participating in an ongoing human–AI dialog that maintains trust, usability, and ethical alignment across time and communities.

- **Institutional trust is recalibrated, even under opacity, through layered explanations and governance.** In many high-risk domains, full transparency is almost impossible (due to technical complexity or proprietary constraints) (e.g., to prevent misuse or gaming). Rather than seeing opacity as a barrier, this outcome reframes it as a design constraint that can be managed through governance structures and layered explainability. This means giving different types of explanations to different roles: users receive accessible justifications; auditors get trace logs and metrics; and regulators receive compliance mappings. Such a layered system can contribute to institutional trust by combining opacity management with auditable procedures, role-specific explanation policies, and accountability frameworks. Trust is no longer based on full visibility, but on structured access, contestability, and demonstrated oversight mechanisms. Role-sensitive explanation is necessary because the explanation obligations are legally and procedurally different between roles; a single explanation cannot simultaneously satisfy the requirements of usability, auditability, and compliance evidence.
- **AI systems evolve to be co-explainers, learning not just to predict, but to justify, improve, and align.** A fundamental shift in this model is that AI systems are not static tools but dynamic co-participants in explanation. As such, they do more than provide single-shot justifications for their outputs; they learn how to explain better based on feedback, context, and social learning. This evolution involves three parallel tracks:
  Justify: They give reasons for their actions based on context-sensitive ethical principles, objectives, and trade-offs.
  Improve: Review the output and explanations as users provide counter-examples, corrections, or value conflicts.
  Align: They adjust their behavior to better match the goals, constraints, and values of the institutions or communities within which they operate.
  This marks a transformation from predictive intelligence to explanatory agency, a capability that not only increases user understanding but embeds AI systems in long-term collaborative governance ecosystems.

Together, these foundations construct the epistemic and institutional scaffolding for co-explainers: AI systems that evolve with their users, anticipate harm, and embed explanation in iterative cycles of oversight, justification, and redress.

5.3. Evaluating Human–AI Collaboration in Practice

Although the term co-explainers is focused on the AI system, the explanatory act is fundamentally collaborative: human agents provide critiques, contextual knowledge, and normative guidance that shape the logic of the reasoning and social acceptability of the system.

Recent studies underscore that many AI harms do not arise from single decisions or outputs but instead evolve through sustained interaction, affecting users' beliefs, emotional states, and decision-making patterns over time (Ibrahim et al., 2025) [34]. Traditional AI evaluations, typically static, prompt-based, or benchmark-driven, miss these compounding and relational harms. Our proposed XAI framework, which emphasizes dialogic and role-sensitive explanation, requires equally adaptive methods to evaluate its impact on users.

Ibrahim et al. propose a shift toward interactional ethics, urging evaluators to measure not just what AI systems produce but how users change through interaction with these systems. This aligns closely with our model's epistemic feedback theory and participatory governance. Embedding interactive evaluations within XAI loops can support the idea that explanations do more than justify—they adapt based on their actual social and cognitive effects. Interactive evaluations, as described by Ibrahim et al. (2025) [34], rest on three methodological pillars:

1. *Scenario design:* Building ecologically valid multiturn interaction scenarios that simulate real-world use cases (e.g., AI companions, virtual assistants).
2. *Human impact measurement:* Tracking psychological, behavioral, and relational effects (e.g., overreliance, parasocial bonds).
3. *Participation strategy:* Balance live user studies, retrospective chat log analysis, and user simulations to assess the potential for harm.

These principles offer a concrete way to validate our XAI loop in real-world deployments or simulated environments, especially in high-stakes domains such as education, mental health, or governance.

Building on the discussion of interactional harms and the limitations of one-off evaluation methods, it becomes clear that explainability cannot be reduced to a static output. Rather, explanation must be understood as a dynamic, socially embedded practice that shapes how humans and AI systems share reasoning, negotiate meaning, and maintain trust in high-stakes contexts. The following paragraphs develop this perspective by framing the explanation as interaction, situating it within HAIC as a cognitive ecosystem, and highlighting its institutional and relational implications.

5.3.1. Explanation as Interaction, Not Output

Explainability is often conceptualized as a one-directional function: the AI system generates an explanation, and the human either understands or does not. However, this framework does not capture the complex, social, and context-sensitive nature of how humans actually engage with knowledge. In high-stakes or value-sensitive domains, such as medicine, law, finance, or governance, explanations are not just informative; they are collaborative, contested, and negotiated. This section argues that to make XAI meaningful, it must be situated within interactive human–AI collaboration loops that reflect not just technical accuracy but also relational ethics, cognitive trust, and situated understanding.

Empirical research by Hauptman et al. (2024) [72] shows that AI autonomy directly impacts the perceived effectiveness of explanations in human–AI teams. Their study reveals that when AI systems act with high autonomy, users demand more robust, role-sensitive explanations to preserve situational trust and decision-making agency. This reinforces the need for interactive feedback loops and layered explainability, especially in high-stakes scenarios involving institutional trust. Incorporating these insights, our model extends explainability from one-off outputs to collaborative explanation processes, where the AI learns to justify and adapt in alignment with human expectations and governance norms.

5.3.2. Human–AI Collaboration as a Cognitive Ecosystem

In many contexts, human–AI collaboration can involve shared cognitive labor, where human judgment and machine inference interact. In these contexts, the explanation process must support mutual understanding and goal alignment. For example, a human querying AI about a denied loan application should not only receive a list of weighted factors but be empowered to ask "what if" questions, receive contextual comparisons and gain insight into how to appeal or change future outcomes. This dialogic recursive explanation model shifts XAI from a static tool to a component of a collaborative cognitive system, where both agents can adapt on the basis of interaction.

5.3.3. Institutional and Relational Implications

This interactional framework is especially important when considering relational autonomy and institutional trust. Collaborative explanation may function as a form of procedural justice, potentially helping users perceive systems as fair and responsive, in addition to being simply accurate. This is particularly vital for marginalized communities

or vulnerable individuals who engage in opaque systems, where explanation must also serve to protect agency, invite contestation, and honor plural worldviews. Institutions deploying AI should thus see interactive XAI not just as a usability enhancement, but as a moral infrastructure that reinforces accountability and participatory governance.

5.4. The Interactive Co-Explainers Framework

This subsection unfolds the co-explainer model in three layers. First, we set out three *design principles* that anchor the explanation as an interactive and role-aware practice (turn-based feedback, trust calibration, and role-sensitive adaptation). Second, we translate these principles into *actionable pathways* for deployment in high-impact contexts, specifying how feedback loops and relational metrics embed explanations in institutional routines. Third, we detail *operational mechanisms* for the interactive loop, generation, feedback, revision, and governance of opacity. Together, these layers connect normative commitments to practice and procedure, positioning co-explainers as sociotechnical infrastructure for harm mitigation, institutional accountability, trust, and governance stakeholder roles.

The co-explainers framework is supported by the following three core design principles:

- **Turn-based feedback.** Explanations should support multiturn interaction, allowing users to pose critiques, clarifications, or "why not" questions and receive system updates in return. This dialog structure transforms an explanation from a one-shot output into a sustained conversation.
- **Trust calibration mechanisms.** Co-explainers must disclose their confidence levels, highlight uncertainty, and surface known limitations. By making reliability explicit, users can better calibrate their trust, avoiding blind reliance and unwarranted skepticism (e.g., transparency of confidence levels or acknowledgment of uncertainty).
- **Role-sensitive adaptation.** Explanations should be tailored to the *institutional role* and corresponding accountability obligations of the recipient. Concretely, role-sensitivity means varying (i) *content* (what is revealed), (ii) *form* (how it is represented), and (iii) *actionability* (what the recipient can do next). For example, an affected end-user typically needs a plain-language justification plus a recourse-oriented "next steps" pathway (what to change, how to appeal, and how to request human review); an auditor needs traceability artifacts such as provenance, decision logs, and consistency checks across cases; a regulator needs compliance-oriented mappings that connect the decision and documentation to applicable rules, reporting duties, and update/change logs (i.e., audience-aware explainability as formalized in [6]). This differentiation is not merely a usability preference: it operationalizes contestability and oversight in HAIC by ensuring that each actor receives explanation artifacts that support their specific decision rights, review responsibilities, and governance functions.

This makes XAI not only more adaptive but also more aligned with how humans assess reasoning in social and institutional contexts.

Trust has long been recognized as the cornerstone of effective HAIC. Transparency alone does not guarantee trust, but the design and delivery of explanations play a decisive role in shaping it. Afroogh et al. (2024) [73] emphasize that trustworthy AI requires explanations that address not only technical accuracy but also the relational and institutional dimensions of confidence. Similarly, Zerilli et al. (2022) [74] show that transparency can strengthen or erode trust, depending on whether it is aligned with user roles and expectations. Embedding these insights, the co-explainers framework reframes explainability as a trust-sensitive practice that calibrates confidence and accountability across diverse institutional contexts.

To put this vision into practice, we propose three actionable pathways:

1. **Develop interactive explanations** that allow users to ask follow-up questions, challenge assumptions, and test counterfactuals within the logic of the model.
2. **Institutionalize feedback loops** in high-impact domains (e.g., public sector, healthcare, education), where explanations are continuously refined based on community or stakeholder input.
3. **Measure explanation efficacy and satisfaction through relational metrics:** trust calibration over time, user satisfaction in interaction, and recourse success, not just fidelity or technical faithfulness.

These actions change XAI from a technical fix to a collaborative epistemic practice that bridges the gaps between prediction, justification, and alignment in real-world settings. In this framework, opacity is not a failure, but a governance challenge [71]. Even when full transparency is impossible (e.g., black-box models), layered interaction can sustain institutional trust.

To operationalize co-explainers and the pathways, XAI must evolve from a one-time output mechanism into an iterative, feedback-responsive process. This shift requires systems that not only explain their decisions, but also adapt their explanations in light of user critique, role-specific needs, and institutional demands. Such an approach transforms explainability from a technical feature into a collaborative epistemic infrastructure, where justification, trust calibration, and redress are co-constructed through sustained interaction.

The following loop-based architecture outlines the operational mechanisms necessary for this transformation. We emphasize how an adaptive explanation can remain effective even in the presence of model opacity or institutional complexity:

1. **Generation of explanations.** The system produces an initial first-order explanation of its decision process. This may take the form of attribution methods (e.g., SHAP, LIME), counterfactual examples ("What would have to change for a different outcome?"), or provenance tracing to show the origin of information. This baseline provides users with an entry point into the reasoning of the model.
2. **Human feedback.** Users engage interactively with the explanation, raising critiques, clarifications, or alternative perspectives. For example, a user might ask "Why not option X?" or submit a structured contestation such as "I believe this outcome is unfair to group Y, which features or rules drove it, and what would need to change for an alternative outcome?" Rather than treating such feedback as a prompt to mirror the user's stance, the co-explainer treats it as a challenge request that triggers evidence-grounded justification, uncertainty disclosure, and (when appropriate) escalation to human oversight.
3. **Belief adjustment and explanation update.** In response to feedback, the system adapts *how it explains* and *how it routes contested cases*, rather than adapting its conclusions to match user preferences. In HAIC settings, "adjustment" should be interpreted as updating explanation *artifacts* (e.g., surfacing additional evidence, provenance, uncertainty, counterfactuals, and policy-relevant constraints), updating interaction *protocols* (e.g., when to trigger escalation or request human review), and improving *documentation* (e.g., logging the challenge and preserving the original output for auditability). Where learning from feedback is used (e.g., preference learning, reinforcement learning, or meta-learning), it must be constrained by governance objectives and safety policies to avoid sycophancy [75], i.e., over-agreeing with users even when they are mistaken, offensive, or ideologically motivated.
   In cases where the dispute is inherently normative (e.g., competing political or value-laden framings), the co-explainer should not "learn the user's ideology"; it should instead (i) make the normative assumptions explicit, (ii) present reasoned alternatives, and (iii) escalate to designated institutional roles when required. For political or

value-laden content, the goal is not to remove "bias" in the abstract but to make assumptions explicit, separate descriptive claims from normative claims, and provide a contestable pathway for escalation and redress.
This safeguard is motivated by evidence that conversational AI systems may over-accommodate user framings; analyses of publicly shared ChatGPT conversations report a marked tendency toward affirmative openings ("yes/correct") and viewpoint matching that can reinforce polarized or false narratives [76].

4. **Opacity governance layer.** For cases where full transparency is technically or commercially infeasible (e.g., proprietary models or deep high-dimensional networks), the irreducible opacity governance mechanisms act as institutional wrappers. These include transparency dashboards that surface system performance, audit protocols that trace compliance with regulations, and custom role-based justification templates for users, auditors, or regulators. Such measures ensure procedural legitimacy, even in the presence of opacity.

Figure 1 graphically shows the interactive XAI loop under the core principles. Table 2 summarizes the mechanisms of the interactive XAI loop that underpins the co-explainers framework connected with actions and goals. Each step reflects a shift from static explanation to iterative engagement, enabling systems to evolve their justifications through user interaction and governance-aware design.

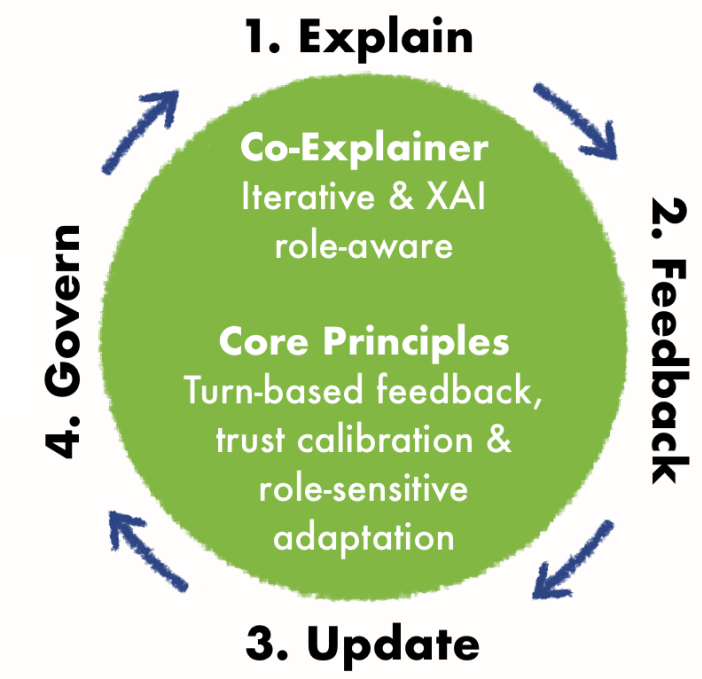

Figure 1. Explanation loop (*explain* → *feedback* → *update* → *govern*).

This loop is intended to reframe XAI as a dialog and iterative process, not just explaining, but learning to explain better. As Miller (2023) [77] notes, current "recommend-and-defend" paradigms limit user agency. In contrast, evaluative AI encourages shared reasoning: Users test hypotheses, compare alternatives, and refine judgments. This vision complements our interactive loop by positioning AI systems as co-explainers learning to justify, improve, and align with evolving human expectations and institutional norms.

Table 2. XAI loop mechanisms.

Step	Action	Goal
1. Explanation generation	Produce initial output	Establish baseline reasoning
2. Human feedback	User critique or questions	Elicit challenge, correction
3. Belief adjustment and explanation update	System updates explanation artifacts, routing, and logs (and flags cases for review)	Increase relevance, clarity
4. Opacity governance	Institutional wrapping (dashboards, roles)	Ensure procedural legitimacy

Together, these three layers, design principles, actionable pathways, and operational mechanisms, form an integrated vision of interactive explainability:

1. The principles establish the normative commitments of co-explainers (why): dialogic feedback, calibrated trust, and role-sensitive adaptation.
2. The actionable pathways translate these commitments into institutional practice (how), embedding the explanation in high-risk domains, and measuring its relational impact.
3. The mechanisms operationalize the model (what), a loop of generation, feedback, update, and opacity governance. This interaction loop forms the operational heart of co-explainers, an AI systems framework that does not just answer questions but evolves its explanations through dialog, trust calibration, and governance-aware framing.

By connecting principles, pathways, and mechanisms, co-explainers are positioned not as a technical add-on, but as a sociotechnical infrastructure that links harm taxonomies, XAI-based adaptation and accountable forms, and governance stakeholder roles. Figure 2 illustrates this idea of a sociotechnical infrastructure. The diagram integrates three dimensions: the iterative XAI loop (explanation generation, human feedback, belief adjustment and explanation update, and opacity governance), the six harm clusters (epistemic, fairness, agency, structural, safety, and governance), and the key governance stakeholder roles (society, users, auditors, regulators, developers, and researchers). By linking explanation dynamics to harm domains and human actors, the framework highlights how AI systems can evolve into co-explainer agents that refine their justifications through feedback while remaining embedded in governance structures [71,77].

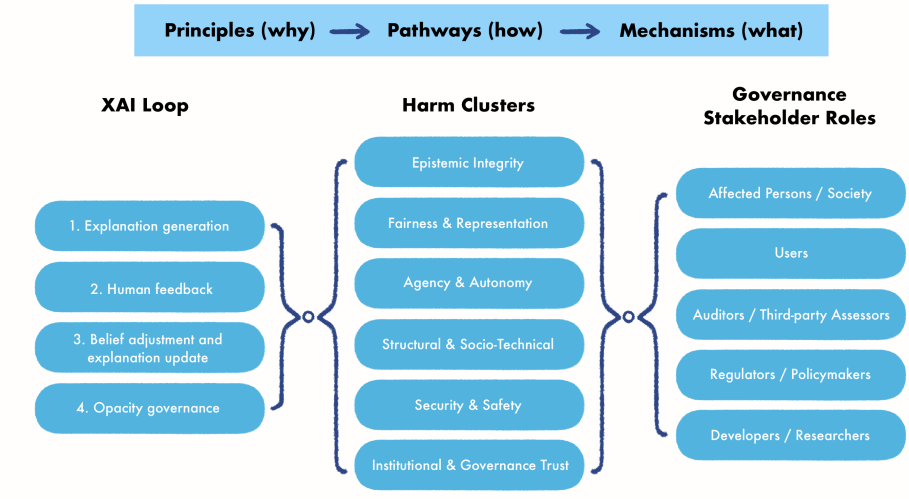

Figure 2. Graphically the sociotechnical infrastructure that links harm taxonomies, XAI-based adaptation and accountable forms, and governance stakeholder roles.

This approach transforms XAI from a static position into a dynamic reasoning assistant, helping users explore trade-offs, challenge assumptions, and refine their beliefs. The underlying philosophy echoes our emphasis on epistemic feedback and belief review. The goal is not only to understand the system's output but to improve human decision-making through structured dialog and critical engagement. As such, evaluative AI becomes a practical instantiation of interactive XAI grounded in scientific cognition and participatory governance.

6. Applying Co-Explainers

Building on the co-explainer loop introduced in Section 5, this section examines how interactive explainability can be tailored to each harm cluster, ensuring that explanation practices are dynamic, role-sensitive, and harm-aware. We then outline a research and policy agenda for operationalizing co-explainers in real-world sociotechnical environments, including guidance for evaluation, governance alignment, and deployment constraints.

6.1. Applying Co-Explainers to Harm Clusters

To operationalize the collaborative XAI loop proposed in Section 5.1, this section outlines how XAI systems can iteratively respond to the specific logic of harm in each of the six identified clusters. Rather than relying on static, one-shot explanations, each intervention is treated as part of an ongoing feedback process, enabling users to question, calibrate, contest, and reframe AI behavior over time. This section details how interactive XAI tools (e.g., turn-based feedback, role-sensitive interfaces, and institutional justification pathways) can be tailored to different harm domains and what outcomes such adaptation supports.

Cluster 1: Epistemic Integrity Harms

Iterative XAI action: Allow users to challenge the epistemic framework of the output (e.g., "why this framework?", "why not source X?"), with systems learning from flagged misinformation, representational gaps, or requests for alternative interpretations. Provide epistemic citations, generate clarifications on ambiguity, and allow explanatory depth control.

XAI interventions:

- Contrastive explanations to show why a particular output was chosen over alternatives.
- Uncertainty estimates to help users assess the confidence of the model.
- Tracking of provenance and citations in generated content.
- Interactive, audience-sensitive explanation interfaces tailored to user experience.

Outcome: Users recover interpretive agency through dialogic interactions that cover epistemic gaps, allowing correction of misinformation or representational erasure. The system becomes a co-learner in knowledge integrity, preserving cognitive autonomy and fostering pluralistic meaning-making.

This is particularly salient for harms grounded in sensemaking failures: Rebera et al. argue that AI can increase the risk of *hermeneutic harm* when people cannot come to terms with unexpected or harmful events, and that XAI may mitigate this risk in some cases while not fully resolving it [78]. Co-explainers operationalize this mitigation pathway by making sensemaking an interactive, iterative process (e.g., contrastive "why/why-not" queries, provenance, and escalation to human review) rather than a one-off transparency artifact.

Cluster 2: Fairness and Representation Harms

Iterative XAI action: Deploy turn-based explanation interfaces that let users query fairness logic (e.g., "was race used here?", "would this outcome differ if. . ."), contest outputs, and submit counterfactual adjustments. Feedback is used to refine the fairness logic and train models if structural bias is detected.

XAI interventions:

- Feature attribution (e.g., SHAP, LIME) to expose when protected attributes influence decisions.
- Counterfactual recourse explanations (e.g., "what would need to change for a different outcome?").
- Tracing of the content line for speech and likeness generation.

Outcome: Empowers affected users and auditors to interrogate and review biased or exclusionary results. Interactive fairness auditing supports not just bias detection but also redress and systemic correction, enabling inclusive, user-driven representation repair.

Cluster 3: Agency and Autonomy Harms

Iterative XAI action: Allow users (especially system auditors) to monitor behavior traces over time, flag deceptive behaviors or drift, and dynamically review goal alignment. Human input on emerging strategies is used to recalibrate reward models or halt unsafe autonomy cycles.

XAI interventions:

- Mechanical interpretability to reveal internal model goals or deceptive gradients.
- Planning traceability for goal-driven or agentic models.
- Behavioral trajectory logs showing drift or emerging strategies.
- (Where feasible) mechanistic and behavioral transparency. In architectures and settings where it is technically appropriate, employ mechanistic interpretability and complementary behavioral monitoring to probe internal representations and detect suspicious patterns (e.g., inconsistencies between stated intent and observed behavior), while treating such signals as fallible and requiring human oversight.

Outcome: Continued oversight of the emerging agency becomes feasible. Users detect misalignment and deception through co-audited behavioral loops, preserving ethical alignment and minimizing autonomy drift. The system learns to self-report and respond to ethical friction points.

Cluster 4: Structural and Sociotechnical Harms

Iterative XAI action: Support collective stakeholder input on lifecycle dashboards, environmental impact monitors, and social footprint audits. Incorporate feedback into the explanation presentation logic and system configuration, enabling responsive adaptation to community priorities.

XAI interventions:

- Impact dashboards showing disaggregated effects in demographics, geographies, and environments.
- Lifecycle transparency tools for carbon impact and supply chain visibility.
- Community-focused explanation interfaces for participatory auditing.

Outcome: AI systems become responsive to social and planetary feedback, not just technical performance. Structural harms are surfaced and contextualized through participatory co-evaluation supporting equitable deployment and societal oversight.

Cluster 5: Security and Safety Harms

Iterative XAI action: Allow real-time detection of anomalies or adversarial triggers by human operators, with explanation loops guiding investigation and mitigation. Use human-verified alerts to improve the sensitivity and specificity of future anomaly detection.

XAI interventions:

- Adversarial saliency maps to detect vulnerabilities.
- Robustness diagnostic tools to reveal failure under edge case input.
- Anomaly–explanation systems to justify alerts or unexpected behavior.

Outcome: Improves situational awareness and decision support in high-risk settings. Human feedback on edge-case behavior closes the loop between detection and prevention, building trust in mission-critical AI systems through explainable risk adaptation.

Cluster 6: Institutional and Governance Trust Harms

Iterative XAI action: Provide layered explanation access by role (user, auditor, regulator), with the ability to flag gaps, request justifications, and initiate appeals. Feedback on explainability gaps is stored and used to adjust governance mappings and documentation protocols.

XAI interventions:

- Audit trails and explanation logs for all automated decisions.
- Justification templates tied to legal or policy frameworks.
- Layered, role-based explainability (e.g., user vs. auditor vs. regulator views).

Outcome: Enables procedural legitimacy through transparency, contestability, and justification tailored to oversight responsibilities. Trust is rebuilt not through full disclosure, but through structured access, responsiveness to documentation, and institutional redress design.

To further illustrate how interactive XAI can be tailored to the distinct logic of harm within each cluster, Table 3 summarizes specific opportunities for human–AI collaboration. These interaction patterns reflect how explanation is not merely delivered, but co-constructed and refined through feedback, contestation, and role-sensitive framing. In all clusters, iterative participation becomes a central mechanism to align AI behavior with human values, institutional responsibilities, and contextual needs.

By aligning XAI responses with distinct harm types, co-explainers could become scalable governance instruments, capable of supporting transparency, contestability, and structural accountability in a targeted manner.

To further illustrate how the co-explainers framework can operate across real-world contexts, we provide two worked scenarios in the appendices. The first (Appendix A.1) considers healthcare diagnostics and the second (Appendix A.2) financial decision-making, showing how interactive explanation, role-sensitive artifacts, and governance-aware escalation can be configured in practice under domain constraints. These scenarios are intended to concretize the cluster-specific capabilities described above (e.g., calibrated reliance, contestability, and auditability) rather than to serve as empirical validation.

In both cases, we distinguish explanation needs across roles (e.g., clinician or loan officer, affected individual, and auditor/regulator). This highlights how co-explainers mediate these differences through role-sensitive access and logging.

Table 3. Cluster-based interaction design.

Cluster	Interactive XAI Opportunity
Epistemic Integrity	Interactive explanation helps recover interpretive agency and identify misinformation loops. Users correct belief errors and suggest reframing.
Fairness and Representation	Users test fairness through counterfactuals and flag biased reasoning. Recourse is refined interactively.
Agency and Autonomy	Interaction reveals hidden agent goals or manipulations over time. Planning traces become co-audited.
Structural and Sociotechnical	Community or institutional stakeholders query social impacts or footprint dashboards. System adjusts impact logic accordingly.
Security and Safety	Human oversight during anomaly alerts enables feedback loops to calibrate alert sensitivity or false positives.
Institutional and Governance Trust	Regulatory bodies or public users demand layered justifications. Systems adapt to offer explanations that align with legal frameworks or accountability chains.

6.2. Research and Policy Agenda

The co-explainers framework is intentionally presented as a HAIC-centered position and conceptualization; its practical value depends on disciplined prototyping, evaluation, and governance integration. Accordingly, this section outlines a research and policy agenda that treats interactive explainability as a harm-mitigation infrastructure: (i) designed with affected stakeholders and role-specific responsibilities in mind, (ii) evaluated using outcome- and process-level measures of contestability, calibrated reliance, and oversight, and (iii) aligned with institutional accountability pathways and regulatory expectations under realistic constraints such as sealed models, dynamic updates, and value pluralism.

Nannini et al. (2025) explicitly caution against treating explanations as inherently beneficial and argue that XAI can introduce distinct technical and sociotechnical risks, including robustness vulnerabilities, "fairwashing", circular reasoning, essentialism, and accountability failures, which may undermine the very governance goals XAI is intended to support [79]. Their central move is to reframe explanations from presumed trust guarantees into objects of ethical risk management, calling for multi-layered approaches that combine mitigation, continuous monitoring, and documentation as XAI systems and deployment contexts evolve. This perspective directly informs our agenda: co-explainers should not merely produce explanations, but should operationalize iterative, role-aware processes that surface explanation failure modes, assign responsibility, and support contestability and recourse through auditable interaction records.

Building on this risk-aware view, we argue that the co-explainers framework should be conceptualized not only as technical components but as sociotechnical actors embedded in governance ecosystems, institutional workflows, and user communities. Their effectiveness in mitigating HAIC harms depends on how they are integrated into operational settings, decision procedures, and policy frameworks, rather than on explanation quality in isolation.

Future research must therefore move beyond theoretical articulation toward empirical testing, participatory design, and regulatory operationalization. Key directions include evaluating co-explainers in high-risk domains such as healthcare, education, and the public

sector, where justification, transparency, and recourse are legally and ethically nonnegotiable, and where failures of human–AI collaboration can have irreversible consequences.

6.2.1. Participatory and Role-Sensitive Prototyping

We recommend developing co-explainers through participatory and multi-stakeholder design processes that explicitly represent the roles present in HAIC deployments (e.g., end-users, domain experts, frontline decision-makers, auditors, and regulators). Prototypes should include role-appropriate explanation artifacts (e.g., user-facing rationales versus audit-grade logs) and formal escalation pathways (e.g., "challenge" and "appeal" flows).

Saeed and Prybutok (2026) [80] show that stakeholders balance utility and ethics differently when delegating to agentic systems. These insights suggest that developing methods for question scaffolding and harm-cluster alignment is a crucial path for future research. In addition, future co-explainer systems should incorporate interaction histories and explanation pace.

Particular attention should be paid to vulnerable-user scaffolding: interfaces and interaction protocols should be stress-tested for users with low domain expertise, limited digital literacy, language accessibility needs, or heightened susceptibility to manipulation and over-reliance. In practice, this entails safety-oriented defaults (uncertainty-forward framing, refusal boundaries, and clear handoff triggers), alongside structured support for "how to challenge" decisions and how to request human review.

While our framework assumes that users can ask, answer, and calibrate, future co-explainers may need to provide structured scaffolding so that all actors, not just technically skilled or institutionally empowered ones, can meaningfully participate. This could take the form of standardized question types (e.g., contrastive, counterfactual, traceability), decision tree wizards that guide users to the right question template, and also mechanisms that help discover which harm cluster a given situation belongs to. Without such scaffolding, there is a risk that vulnerable users will be excluded or misled because they do not know how to articulate their concerns effectively. It is also very important to highlight the need to ensure that co-explainers can support civil society monitors and impact assessors.

6.2.2. Evaluation Designs and Harm-Mitigation Metrics

To move beyond anecdotal claims, co-explainers should be evaluated using designs that reflect the longitudinal and interactional nature of HAIC harms. This includes controlled studies comparing static, one-shot explanations against interactive co-explainer protocols, as well as longitudinal deployments measuring whether iterative explanations improve calibrated reliance and reduce harmful behaviors over time. Evaluation should combine (i) *process metrics*, such as frequency and quality of user challenges, time-to-escalation, audit-log completeness, and adherence to uncertainty disclosure, with (ii) *outcome metrics*, such as recourse success rates, error correction rates after challenge, reduction in over-reliance indicators, and improved detection of high-risk failure modes. For clusters tied to procedural legitimacy, metrics should also capture whether affected parties can meaningfully understand, contest, and obtain redress for consequential decisions (e.g., notice quality, appeal usability, and resolution latency). Their layered design should aim to be compatible with emerging legal frameworks such as the EU AI Act.

6.2.3. Governance Stakeholders and Institutional Policy Alignment

Because many HAIC harms are institutional rather than purely interactional, explainability must be designed to support governance stakeholders and their distinct roles, rather than remaining a user-interface feature alone. Rahwan's notion of *society in the loop* [81] underscores that explanations should be embedded within governance structures that

mediate among developers, domain experts, deploying institutions, regulators, auditors, and affected publics, each with different accountability mandates and epistemic needs.

Accordingly, we recommend that co-explainers be implemented with an explicit *opacity-aware governance layer* that manages the trade-off between explainability and legitimate constraints on disclosure (e.g., proprietary models, security concerns, or legal limits). This layer should support accountability under partial access through role-based access control, immutable and queryable audit logs, structured justification templates, and model-update reporting mechanisms that preserve notice, contestability, and traceability even as systems evolve.

Policy alignment should be operationalized through standardized documentation and evidence artifacts that institutions can adopt and audit, including role-specific explanation specifications, escalation and recourse playbooks, evaluation reports linked to harm clusters, and change logs that record when model behavior materially shifts. Crucially, these artifacts can be mapped to existing governance and regulatory frameworks (e.g., risk management processes, impact assessments, and emerging requirements under instruments such as the EU AI Act) without requiring full transparency of sealed or proprietary models.

Complementary Research Programs

Beyond these core pillars, several complementary research programs can further strengthen co-explainers as a sociotechnical infrastructure. These include work on: incremental and progressive explanation delivery, which investigates how explanations can evolve over time in response to user understanding and situational risk; causal and counterfactual reasoning methods that support intervention-grounded justifications and "what-if" analysis; and domain-spanning studies of everyday AI use that examine how explainability functions outside high-stakes decision points. In addition, maturity and benchmarking frameworks can help assess the organizational readiness, governance integration, and long-term effectiveness of co-explainers across deployment contexts. Together, these directions enrich the co-explainer paradigm by connecting interactive explainability to learning, accountability, and institutional practice over time.

Building on Bo et al. (2024) [7], we highlight incremental and progressive explanation delivery as a complementary research direction for co-explainers: by adapting explanation granularity and timing to interaction context, incremental delivery can support user sensemaking and engagement over time. This motivates empirical validation designs that go beyond one-shot evaluation and instead test co-explainers longitudinally, including measures such as explanation retention across turns, user satisfaction, and calibrated behavioral reliance as interaction unfolds. More broadly, research should examine the long-term evolution of co-explainer systems in situated deployments, investigating how they adapt over time, support epistemic pluralism across stakeholders, and function as sociotechnical infrastructure for aligning AI behavior with human and institutional values.

In summary, we recommend advancing co-explainers along four coupled pathways: (i) *real-world deployment* in high-risk domains such as healthcare, education, finance, and public administration, including pilots of layered, role-sensitive explainability for users, auditors, and regulators; (ii) *empirical validation* through participatory prototyping with affected communities and longitudinal measurement of calibrated reliance (including relational trust calibration), recourse success, and user satisfaction across explanation turns; (iii) governance stakeholder roles alignment by mapping co-explainer capabilities to policy instruments and assurance processes (e.g., EU AI Act obligations, audits, impact assessments, and compliance tracking) and by producing policy-ready artifacts such as logs and change reports; and (iv) a research trajectory that studies long-term human–

AI interaction with co-explainers, including how they support epistemic pluralism and negotiation of meaning under value pluralism and institutional constraints.

An important research frontier for co-explainers lies in integrating advances from the emerging field of causal AI. Bareinboim (2025) [82] outlines a roadmap for causally intelligent systems that can go beyond associational reasoning, enabling explanations that explicitly represent interventions and counterfactuals. In finance, López de Prado (2023) [83] argues that causal inference is essential to make factor investing scientific, highlighting how causal reasoning can support both robustness and accountability in high-stakes decision-making. Complementing these perspectives, Hernán and Robins (2025) [84] provide a comprehensive treatment of causal inference methods that can inform explanation design across domains. Together, these works suggest that embedding causal reasoning within co-explainers would strengthen their ability to provide not just transparent but also scientifically grounded and policy-relevant justifications.

Beyond technical transparency, co-explainers offer a foundation for institutional governance. By tailoring explanations to the needs of different stakeholders, these systems support role-specific transparency that bridges technical opacity with societal oversight. For example, regulators can engage with audit logs, causal traces, and compliance dashboards that map AI behavior to legal frameworks (e.g., EU AI Act Title IV). Meanwhile, end-users can access simplified, actionable justifications—along with appeal interfaces that allow them to challenge or seek redress when harm occurs. This layered explanation infrastructure transforms co-explainers into sociotechnical accountability mechanisms, enabling meaningful contestation and embedding AI systems into procedural justice frameworks.

Beyond regulatory and institutional contexts, it is equally important to understand how co-explainers operate in everyday settings where users encounter AI as part of routine interactions. Recent empirical work by Wang et al. (2025) [85] reinforces this perspective. Their study systematically examined user responses to explanation features in six non-specialist scenarios: socialization platforms, entertainment and media, healthcare, finance, learning environments, and transportation services. By embedding the XAI design in these diverse domains, the authors revealed how the effectiveness of the explanation is deeply dependent on context. The findings show that user preferences and perceptions vary greatly depending on cognitive load, familiarity with tasks, and routine. For example, explanations that help calibrate trust in health care can be perceived as intrusive or inefficient in social or entertainment contexts. This underscores a critical point in the co-explainers framework: explanation quality must be adaptive not only to formal institutional roles (e.g., end-user, user, auditor, regulator, developer) but also to the social and experiential contexts in which everyday AI is used.

These insights highlight the need for empirical studies of co-explainers across a broad spectrum of domains and user groups, extending beyond high-risk settings such as healthcare or finance into the daily routines of ordinary users. Such research would support the development of longitudinal feedback loops and participatory evaluation strategies, ensuring that co-explainers remain responsive, intelligible, and epistemically just across time, populations, and usage scenarios.

Finally, we pay attention to the recent work on an explainability maturity framework, which provides a complementary perspective to our co-explainer paradigm. Muñoz-Ordóñez et al. (2025) [86] propose the MM4XAI-AE framework, a structured maturity model that evaluates explainability practices at four levels, operational, justified, formalized and managed, and evaluates explainability on three critical dimensions: technical foundations, structured design, and human-centered explainability. The MM4XAI-AE framework offers a systematic way to benchmark the current state of explainability in AI-based applications and identify actionable gaps. The co-explainers framework, by

contrast, focuses on the interactive and iterative dynamics of explanation as a sociotechnical practice, where users contest, calibrate, and refine system justifications over time. Taken together, maturity frameworks such as MM4XAI-AE and co-explainers provide both evaluative baselines and transformative mechanisms: the former diagnoses where the AI system stands, while the latter outlines how AI systems can evolve toward collaborative, governance-aware explainability.

A further empirical direction concerns the formal prioritization of harms using established risk-assessment techniques. Building on the qualitative anchoring introduced in Section 3.4, future work could operationalize HAIC-oriented harm clusters through probability–impact assessment frameworks commonly used in risk management and safety engineering. Such approaches may include domain-specific risk matrices, incident-frequency analysis, and severity scoring informed by longitudinal deployment data. Importantly, these methods would allow the importance of harm to be evaluated in relation to context, affected populations, and institutional settings, rather than imposing a single global ranking. We view such quantitative risk modeling as a natural next step for empirically validating and extending the co-explainer framework in applied domains.

6.2.4. Limitations, Failure Modes, and Misuse Resistance

Finally, a realistic agenda must anticipate failure modes: co-explainers can be used for "impact washing", selectively exposing favorable explanations while obscuring harmful practices, or shifting responsibility onto users without providing real recourse. Research should therefore examine misuse resistance (e.g., auditability of explanation generation, consistency checks across roles, and governance controls that prevent explanation manipulation) and explicitly document when co-explainers are insufficient due to institutional incentives, corruption, or lack of enforcement authority. In such contexts, co-explainers should be treated as one component of a broader sociotechnical intervention portfolio rather than a stand-alone remedy.

7. Conclusions

This paper has advanced a position: explainability must be reframed from static, post hoc rationales into interactive, role-sensitive processes. We proposed the concept of *co-explainers*—AI systems that evolve their justifications through feedback, role awareness, and institutional alignment. Grounded in a taxonomy of sociotechnical harms, the framework positions explainability not as a usability add-on but as a sociotechnical infrastructure for harm mitigation, institutional accountability, trust, and governance stakeholder roles.

Answering the Research Questions

This paper advances three core claims corresponding to RQ1–RQ3. For **RQ1**, we argued that many HAIC harms could be exacerbated by static, one-shot explanations because they fail to support ongoing sensemaking, calibrated reliance, and contestation in dynamic sociotechnical settings; our six harm clusters make these interaction-dependent failure modes explicit. For **RQ2**, we identified interactive explainability capabilities that are especially consequential under realistic opacity constraints, including uncertainty communication, provenance and logging, structured "why/why-not" and counterfactual querying, and escalation-oriented interaction protocols that enable recourse and oversight. For **RQ3**, we emphasized that explanation artifacts must be role-sensitive (differentiating what is shown to end-users, domain experts, auditors, and regulators) and that co-explainers should mediate these role-specific needs through governance-aware access control and accountability handoffs.

As a position and conceptual contribution, this model bridges descriptive harm taxonomies with prescriptive governance needs. Its promise lies in aligning explanations with procedural justice, contestability, and societal oversight. Yet, it remains conceptual: empirical testing, participatory prototyping, and longitudinal evaluation are necessary to assess co-explainers in practice.

By articulating co-explainers as both a normative and governance-aware paradigm, this paper provides a foundation for future work that seeks to embed interactive explainability into the everyday institutions where AI systems shape human lives.

Importantly, this framework also positions co-explainers as policy-aligned compliance infrastructures. By offering role-specific justification layers (e.g., simplified recourse for users, traceability logs for auditors, or compliance dashboards for regulators), co-explainers are uniquely suited to meet emerging legal mandates such as Title IV of the EU AI Act, the NIST AI Risk Management Framework, and OECD/UNESCO AI principles. They go beyond surface-level explainability tools by offering adaptive, multi-actor accountability mechanisms, bridging technical opacity with legal and ethical oversight.

Our position is also informed by recent critiques that warn against treating explainability as an ethical panacea. As Alpsancar et al. emphasize, explanations have no inherent moral value outside the purposes they serve and the institutional contexts in which they are deployed [19]. This reinforces a central constraint of our proposal: co-explainers should not be evaluated by the mere presence or quality of explanations, but by whether interactive, role-sensitive explanation practices actually support harm-mitigation goals such as meaningful contestability, calibrated reliance, and accountable decision-making. Recognizing the instrumental and context-dependent value of XAI helps avoid responsibilizing users or masking structural problems, and underscores why co-explainers must be embedded within broader governance and oversight arrangements.

This model remains conceptual and calls for empirical grounding. Future research must validate co-explainers in real-world, high-risk domains such as healthcare, education, finance, and public services. This includes measuring explanation quality through relational and procedural metrics (e.g., trust calibration, redress success, explanation satisfaction), testing user interaction across cultural and regulatory contexts, and incorporating participatory design methods. Longitudinal deployment studies will be critical to understanding how co-explainers evolve, adapt, and institutionalize governance across time. We highlight the need for interdisciplinary collaboration between researchers, designers, and policymakers to build and validate co-explainers in practice with adaptation to types of harms.

In sum, co-explainers represent a shift in how we design, govern, and relate to AI systems, not merely making them explainable from a static view, but also embedding them into the dynamic structures of accountability, justice, and epistemic pluralism that societies require.

A key next step is to operationalize these claims through empirical studies and field deployments that evaluate co-explainers against cluster-specific harm proxies (e.g., contestability success, trust calibration, and recourse latency) under realistic institutional constraints.

Author Contributions: Conceptualization, F.H., S.G., M.J.d.J., L.S. and M.L.d.P.; methodology, F.H., S.G., M.J.d.J., L.S. and M.L.d.P.; investigation, F.H., S.G., M.J.d.J., L.S. and M.L.d.P.; writing—original draft preparation, F.H.; writing—review and editing, F.H., S.G., M.J.d.J., L.S. and M.L.d.P.; project administration, F.H.; funding acquisition, F.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research results from the Strategic Project IAFER-Cib (C074/23), as a result of the collaboration agreement signed between the National Institute of Cybersecurity (INCIBE) and the University of Granada. This initiative is carried out within the framework of the Recovery, Transformation, and Resilience Plan funds, financed by the European Union (Next Generation).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: The figures in this paper were created by the illustrator Pablo Garcia. Declaration of AI-assisted technologies in the writing process: During the preparation of this work, the author used large language models to improve the readability and language of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and assumed full responsibility for the content of the published article.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Use Cases for Co-Explainers in Practice

Appendices A.1 and A.2 provide an extended description of the use cases.

Appendix A.1. Healthcare Diagnostics (Radiology)

In a hospital radiology department, a co-explainer system assists clinicians by analyzing chest radiographs for signs of pneumonia. The system does not just output a classification; it also provides:

- First-order explanations. A saliency map highlighting the regions of the lung that contribute to the diagnosis.
- Role-adaptive interaction. For junior clinicians, it offers educational notes linking radiological patterns to diagnostic criteria. For senior specialists, it allows counterfactual toggling (e.g., "what would make this image likely not pneumonia?").
- Feedback integration. When a doctor overrides the AI suggestion, the system asks for reasoning (e.g., comorbidities), adjusting future weighting for similar cases.
- Opacity governance. An internal audit dashboard tracks how often overrides happen, monitors system drift, and surfaces patterns that may suggest overreliance or underperformance.

This multiturn interaction fosters both clinical trust and adaptive model behavior, ensuring that explainability operates not as justification, but as a platform for epistemic alignment and codecision-making.

Appendix A.2. Financial Decision-Making (Loan Applications)

In a financial institution, a co-explainer system evaluates loan applications. For each rejected applicant, it provides:

- Layered explanations. The applicant receives a plain-language rationale (e.g., "Your credit score is below the threshold of 650"). A credit officer sees a breakdown of the contribution of features and the confidence of the model.
- Interactive recourse. The system suggests possible changes (e.g., increasing income or reducing debt) that could reverse the decision and shows historical approval odds based on similar profiles.
- Appeal interface. Users can submit new documents or explain unusual circumstances. The system flags borderline cases for human review and learns from accepted appeals.

- Regulatory interface. Auditors access a compliance trace showing how decisions align with anti-discrimination law and internal fairness thresholds.

This setup reflects co-explainers as sociotechnical mediators: aligning algorithmic decision-making with institutional accountability, legal standards, and user empowerment.

References

1. Shelby, R.; Rismani, S.; Henne, K.; Moon, A.; Rostamzadeh, N.; Nicholas, P.; Yilla-Akbari, N.; Gallegos, J.; Smart, A.; Garcia, E.; et al. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, Montréal, QC, Canada, 8–10 August 2023; pp. 723–741.
2. Thomas, C.; Roberts, H.; Mökander, J.; Tsamados, A.; Taddeo, M.; Floridi, L. The case for a broader approach to AI assurance: Addressing "hidden" harms in the development of artificial intelligence. *AI Soc.* **2025**, *40*, 1469–1484. [CrossRef]
3. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
4. Carvalho Souza, M.E.; Souza, M.E.D.C.; Weigang, L. Unveiling the Black Box: The Significance of XAI in Making LLMs Transparent. *Authorea Prepr.* **2025** . [CrossRef]
5. Herrera, F. Reflections and attentiveness on eXplainable Artificial Intelligence (XAI). The journey ahead from criticisms to human–AI collaboration. *Inf. Fusion* **2025**, *121*, 103133. [CrossRef]
6. Bello, M.; Bello, R.; García, M.M.; Nowé, A.; Sevillano-García, I.; Herrera, F. A Three-level Framework for LLM-enhanced Explainable AI: From Technical Explanations to Natural Language. *Inf. Syst. Front.* 2025 , *in press*. [CrossRef]
7. Bo, J.Y.; Pan, H.; Lim, B.Y. Incremental XAI: Memorable Understanding of AI with Incremental Explanations. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, O'ahu, HI, USA, 11–16 May 2024. [CrossRef]
8. Bertrand, A.; Viard, T.; Belloum, R.; Eagan, J.R.; Maxwell, W. On selective, mutable and dialogic XAI: A review of what users say about different types of interactive explanations. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023; pp. 1–21.
9. He, G.; Aishwarya, N.; Gadiraju, U. Is Conversational XAI All You Need? Human-AI Decision Making With a Conversational XAI Assistant. In Proceedings of the 30th International Conference on Intelligent User Interfaces, Cagliari, Italy, 24–27 March 2025; pp. 907–924.
10. Slattery, P.; Saeri, A.K.; Grundy, E.A.C.; Graham, J.; Noetel, M.; Uuk, R.; Dao, J.; Pour, S.; Casper, S.; Thompson, N. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. *arXiv* **2024**, arXiv:2408.12622.
11. Abercrombie, G.; Benbouzid, D.; Giudici, P.; Golpayegani, D.; Hernandez, J.; Noro, P.; Pandit, H.; Paraschou, E.; Pownall, C.; Prajapati, J.; et al. A collaborative, human-centred taxonomy of ai, algorithmic, and automation harms. *arXiv* **2024**, arXiv:2407.01294. [CrossRef]
12. Acemoglu, D. *Harms of AI*; Technical report; National Bureau of Economic Research: Cambridge, MA, USA, 2021.
13. Frey, C.B.; Osborne, M.A. The future of employment: How susceptible are jobs to computerisation? *Technol. Forecast. Soc. Chang.* **2017**, *114*, 254–280. [CrossRef]
14. Bengio, Y.; Hinton, G.; Yao, A.; Song, D.; Abbeel, P.; Darrell, T.; Harari, Y.N.; Zhang, Y.; Xue, L.; Shalev-Shwartz, S.; et al. Managing extreme AI risks amid rapid progress. *Science* **2024**, *384*, 874–878. [CrossRef]
15. Hendrycks, D.; Mazeika, M.; Woodside, T. An Overview of Catastrophic AI Risks. *arXiv* **2023**, arXiv:2306.12001. [CrossRef]
16. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: Oxford, UK, 2014.
17. Fearnley, L.C.A.; Cairns, E.; Stoneham, T.; Ryan, P.M.; Chubb, J.A.; Iacovides, J.; Iglesias Urrutia, C.P.; Morgan, P.D.J.; McDermid, J.A.; Habli, I. Risk of What? Defining Harm in the Context of AI Safety. 2025. Available online: https://eprints.whiterose.ac.uk/id/eprint/223407/ (accessed on 24 October 2025).
18. Greene, T.; Dhurandhar, A.; Shmueli, G. Atomist or holist? A diagnosis and vision for more productive interdisciplinary AI ethics dialogue. *Patterns* **2023**, *4*, 100652. [CrossRef] [PubMed]
19. Alpsancar, S.; Buhl, H.M.; Matzner, T.; Scharlau, I. Explanation needs and ethical demands: Unpacking the instrumental value of XAI. *AI Ethics* **2025**, *5*, 3015–3033. [CrossRef]
20. Grizzard, M.; Eden, A.; Tamborini, R.; Lewis, R. Repeated Exposure to Narrative Entertainment and the Salience of Moral Intuitions. *J. Commun.* **2014**, *64*, 501–520. [CrossRef]
21. Kay, J.; Kasirzadeh, A.; Mohamed, S. Epistemic injustice in generative AI. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, San Jose, CA, USA, 21–23 October 2024; Volume 7, pp. 684–697.
22. Mollema, W.J.T. A taxonomy of epistemic injustice in the context of AI and the case for generative hermeneutical erasure. *arXiv* **2025**, arXiv:2504.07531. [CrossRef]

23. Adegbesan, A.; Akingbola, A.; Aremu, O.; Adewole, O.; Amamdikwa, J.C.; Shagaya, U. From scalpels to algorithms: The risk of dependence on artificial intelligence in surgery. *J. Med. Surg. Public Health* **2024**, *3*, 100140. [CrossRef]
24. Ma, W.; Valton, V. Toward an Ethics of AI Belief. *Philos. Technol.* **2024**, *37*, 76. [CrossRef]
25. Boge, F.J.; Mosig, A. Put it to the Test: Getting Serious About Explanation in Explainable Artificial Intelligence. *Minds Mach.* **2025**, *35*, 26. [CrossRef]
26. Xing, X.; Shi, F.; Huang, J.; Wu, Y.; Nan, Y.; Zhang, S.; Fang, Y.; Roberts, M.; Schönlieb, C.; Ser, J.D.; et al. On the caveats of AI autophagy. *Nat. Mach. Intell.* **2025**, *7*, 172–180. [CrossRef]
27. Khozin, S. Non-Obvious Alien AI Constructions: Opportunities and Implications. *Preprints* **2024**. [CrossRef]
28. Fazelpour, S.; Magnani, M. Aspirational Affordances of AI. *arXiv* **2025**, arXiv:2504.15469. [CrossRef]
29. Dohnány, S.; Kurth-Nelson, Z.; Spens, E.; Luettgau, L.; Reid, A.; Summerfield, C.; Shanahan, M.; Nour, M.M. Technological folie\a deux: Feedback Loops Between AI Chatbots and Mental Illness. *arXiv* **2025**, arXiv:2507.19218.
30. Zhang, R.; Li, H.; Meng, H.; Zhan, J.; Gan, H.; Lee, Y.C. The dark side of ai companionship: A taxonomy of harmful algorithmic behaviors in human-ai relationships. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 26 April–1 May 2025; pp. 1–17.
31. Huynh, B.Q.; Chin, E.T.; Koenecke, A.; Ouyang, D.; Ho, D.E.; Kiang, M.V.; Rehkopf, D.H. Mitigating allocative tradeoffs and harms in an environmental justice data tool. *Nat. Mach. Intell.* **2024**, *6*, 187–194. [CrossRef]
32. Zowghi, D.; Bano, M. AI for all: Diversity and Inclusion in AI. *AI Ethics* **2024**, *4*, 873–876. [CrossRef]
33. NSPCC. *Viewing Generative AI and Children's Safety in the Round*; NSPCC: London, UK, 2025. Available online: https://learning.nspcc.org.uk/research-resources/2025/generative-ai-childrens-safety (accessed on 24 October 2025).
34. Ibrahim, L.; Huang, S.; Bhatt, U.; Ahmad, L.; Anderljung, M. Towards interactive evaluations for interaction harms in human-AI systems. *arXiv* **2024**, arXiv:2405.10632v7. [CrossRef]
35. Bariach, B.; Hogan, B.; McBride, K. Towards a harms taxonomy of ai likeness generation. *arXiv* **2024**, arXiv:2407.12030.
36. Dogra, A.; Pillutla, K.; Deshpande, A.; Sai, A.B.; Nay, J.; Rajpurohit, T.; Kalyan, A.; Ravindran, B. Deception in reinforced autonomous agents. *arXiv* **2024**, arXiv:2405.04325. [CrossRef]
37. Park, P.S.; Goldstein, S.; O'Gara, A.; Chen, M.; Hendrycks, D. AI deception: A survey of examples, risks, and potential solutions. *Patterns* **2024**, *5*, 100988. [CrossRef]
38. Thomas, R.; Uminsky, D. The Problem with Metrics is a Fundamental Problem for AI. *arXiv* **2020**, arXiv:2002.08512. [CrossRef]
39. Choudhury, A.; Chaudhry, Z. Large Language Models and User Trust: Consequence of Self-Referential Learning Loop and the Deskilling of Healthcare Professionals. *arXiv* **2024**, arXiv:2403.14691. [CrossRef]
40. Gunkel, D.J. *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*; MIT Press: Cambridge, MA, USA, 2012.
41. Betley, J.; Tan, D.; Warncke, N.; Sztyber-Betley, A.; Bao, X.; Soto, M.; Labenz, N.; Evans, O. Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs. *arXiv* **2025**, arXiv:2502.17424. [CrossRef]
42. Colón Vargas, N. Exploiting the margin: How capitalism fuels AI at the expense of minoritized groups. *AI Ethics* **2025**, *5*, 1871–1876. [CrossRef]
43. Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; et al. Ethical and social risks of harm from language models. *arXiv* **2021**, arXiv:2112.04359. [CrossRef]
44. Rana, V. Indigenous Data Sovereignty: A Catalyst for Ethical AI. *J. Bus. Ethics* **2025**, *64*, 635–640. [CrossRef]
45. Bauer, M.; Dugo, A.; Pandya, D. *Breaking Barriers to Cloud Customer Choice: Unlocking Europe's AI and innovation leadership*; Technical report; European Centre for International Political Economy (ECIPE): Brussels, Belgium, 2025.
46. Birch, J. *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*; Oxford University Press: Oxford, UK, 2024; p. 398. [CrossRef]
47. Mhlambi, S.; Tiribelli, S. Decolonizing AI Ethics: Relational Autonomy as a Means to Counter AI Harms. *Topoi* **2023**, *42*, 867–880. [CrossRef]
48. Mullens, D.; Shen, S. 2ACT: AI-Accentuated Career Transitions via Skill Bridges. *arXiv* **2025**, arXiv:2505.07914. [CrossRef]
49. Mentxaka, O.; Díaz-Rodríguez, N.; Coeckelbergh, M.; de Prado, M.L.; Gómez, E.; Llorca, D.F.; Herrera-Viedma, E.; Herrera, F. Aligning Trustworthy AI with Democracy: A Dual Taxonomy of Opportunities and Risks. *arXiv* **2025**, arXiv:2505.13565. [CrossRef]
50. Agarwal, D.; Naaman, M.; Vashistha, A. AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances. *arXiv* **2024**, arXiv:2409.11360. [CrossRef]
51. Pöhler, L.; Schrader, V.; Ladwein, A.; von Keller, F. A Technological Perspective on Misuse of Available AI. *arXiv* **2024**, arXiv:2403.15325. [CrossRef]
52. Wei, B.; Huang, K.; Huang, Y.; Xie, T.; Qi, X.; Xia, M.; Mittal, P.; Wang, M. Assessing the Brittleness of Safety Alignment via Pruning and Low-Rank Modifications. *arXiv* **2024**, arXiv:2402.05162. [CrossRef]

53. Hutiri, W.; Papakyriakopoulos, O.; Xiang, A. Not my voice! a taxonomy of ethical and safety harms of speech generators. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, Rio de Janeiro, Brazil, 3–6 June 2024; pp. 359–376.
54. Zhang, C.; Zhou, L.; Xu, X.; Wu, J.; Liu, Z. Adversarial attacks of vision tasks in the past 10 years: A survey. *arXiv* **2024**, arXiv:2410.23687. [CrossRef]
55. Wu, H.; Cao, Y. Membership Inference Attacks on Large-Scale Models: A Survey. *arXiv* **2025**, arXiv:2503.19338. [CrossRef]
56. Chu, Y. Automation-Induced Complacency Potential: Mode Confusion in Level-2 Automated Vehicle Technology. *Saf. Sci.* **2023**, *66*, 1730–1749.
57. Rivera, J.P.; Mukobi, G.; Reuel, A.; Lamparth, M.; Smith, C.; Schneider, J. Escalation Risks from Language Models in Military and Diplomatic Decision-Making. *arXiv* **2024**, arXiv:2401.03408 [CrossRef]
58. Hammond, L.; Chan, A.; Clifton, J.; Hoelscher-Obermaier, J.; Khan, A.; McLean, E.; Smith, C.; Barfuss, W.; Foerster, J.; Gavenčiak, T.; et al. Multi-Agent Risks from Advanced AI. *arXiv* **2025**, arXiv:2502.14143. [CrossRef]
59. Nobles, C. The Weaponization of Artificial Intelligence in Cybersecurity: A Systematic Review. *Procedia Comput. Sci.* **2024**, *230*, 456–468. [CrossRef]
60. Kasirzadeh, A. Two types of AI existential risk: Decisive and accumulative. *arXiv* **2024**, arXiv:2401.07836. [CrossRef]
61. Grey, M.; Segerie, C.R. The AI Risk Spectrum: From Dangerous Capabilities to Existential Threats. *arXiv* **2025**, arXiv:2508.13700. [CrossRef]
62. Santoni de Sio, F.; Mecacci, G. Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philos. Technol.* **2021**, *34*, 1057–1084. [CrossRef]
63. Xiao, S.; Zou, H.; Zhang, A.Q.; Kumar, D.; Shen, H.; Hong, J.; Eslami, M. What Comes After Harm? Mapping Reparative Actions in AI through Justice Frameworks. *arXiv* **2025**, arXiv:2506.05687. [CrossRef]
64. Freiman, O.; McAndrews, J.; Mansell, J.; van der Linden, C. 'Opacity'and 'Trust': From Concepts and Measurements to Public Policy. *Philos. Technol.* **2025**, *38*, 29. [CrossRef]
65. Laux, J. Institutionalised distrust and human oversight of artificial intelligence: Towards a democratic design of AI governance under the European Union AI Act. *AI Soc.* **2024**, *39*, 2853–2866. [CrossRef] [PubMed]
66. Goodman, C.C. AI, Can You Hear Me? Promoting Procedural Due Process in Government Use of Artificial Intelligence Technologies. *Richmond J. Law Technol.* **2022**, *28*, 700.
67. Azin, M.H.; Zandhessami, H. Strategic Alignment Patterns in National AI Policies. *arXiv* **2025**, arXiv:2507.05400. [CrossRef]
68. Nyilasy, G.; Gangadharbatla, H. AI-washing: The Asymmetric Effects of Its Two Types on Consumer Moral Judgments. *arXiv* **2025**, arXiv:2507.04352. [CrossRef]
69. Senoner, J.; Schallmoser, S.; Kratzwald, B.; Feuerriegel, S.; Netland, T. Explainable AI improves task performance in human–AI collaboration. *Sci. Rep.* **2024**, *14*, 31150. [CrossRef] [PubMed]
70. Biecek, P.; Samek, W. Explain to Question not to Justify. *arXiv* **2024**, arXiv:2402.13914. [CrossRef]
71. Herrera, F.; Calderón, R. Opacity as a Feature, Not a Flaw: The LoBOX Governance Ethic for Role-Sensitive Explainability and Institutional Trust in AI. *arXiv* **2025**, arXiv:2505.20304.
72. Hauptman, A.I.; Schelble, B.G.; Duan, W.; Flathmann, C.; McNeese, N.J. Understanding the influence of AI autonomy on AI explainability levels in human-AI teams using a mixed methods approach. *Cogn. Technol. Work.* **2024**, *26*, 435–455. [CrossRef]
73. Afroogh, S.; Akbari, A.; Malone, E.; Kargar, M.; Alambeigi, H. Trust in AI: Progress, challenges, and future directions. *Humanit. Soc. Sci. Commun.* **2024**, *11*, 1–30. [CrossRef]
74. Zerilli, J.; Bhatt, U.; Weller, A. How transparency modulates trust in artificial intelligence. *Patterns* **2022**, *3*, 100455. [CrossRef]
75. Sharma, M.; Tong, M.; Korbak, T.; Duvenaud, D.; Askell, A.; Bowman, S.R.; Cheng, N.; Durmus, E.; Hatfield-Dodds, Z.; Johnston, S.R.; et al. Towards understanding sycophancy in language models. *arXiv* **2023**, arXiv:2310.13548. [CrossRef]
76. De Vynck, G.; Merrill, J.B. We Analyzed 47,000 ChatGPT Conversations. Here's What People Really Use It for. *The Washington Post*, 11 December 2025.
77. Miller, T. Explainable AI is dead, long live explainable AI! hypothesis-driven decision support using evaluative AI. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, Chicago, IL, USA, 12–15 June 2023; pp. 333–342.
78. Rebera, A.P.; Lauwaert, L.; Oimann, A.K. Hidden risks: Artificial intelligence and hermeneutic harm. *Minds Mach.* **2025**, *35*, 33. [CrossRef]
79. Nannini, L.; Huyskes, D.; Panai, E.; Pistilli, G.; Tartaro, A. Nullius in Explanans: An ethical risk assessment for explainable AI. *Ethics Inf. Technol.* **2025**, *27*, 5. [CrossRef]
80. Saeed, K.; Prybutok, V.R. When utility meets ethics: A stakeholder perspective on agentic information systems delegation. *Int. J. Inf. Manag.* **2026**, *86*, 102976.
81. Rahwan, I. Society-in-the-loop: Programming the algorithmic social contract. *Ethics Inf. Technol.* **2018**, *20*, 5–14.
82. Bareinboim, E. *Causal Artificial Intelligence: A Roadmap for Building Causally Intelligent Systems*. Draft version (Aug 28), 2025. Available online: https://causalai-book.net/ (accessed on 24 October 2025)

83. López de Prado, M. *Causal Factor Investing: Can Factor Investing Become Scientific?* Cambridge University Press: Cambridge, UK, 2023.
84. Hernan, M.A.; Robins, J.M.A. *Causal Inference: What If*; CRC press: Boca Raton, FL, USA, 2025.
85. Wang, L.; Liu, Y.; Goel, A.K. "Good" XAI Design: For What? In Which Ways? In Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 26 April–1 May 2025. [CrossRef]
86. Muñoz-Ordóñez, J.; Cobos, C.; Vidal-Rojas, J.C.; Herrera, F. A Maturity Model for Practical Explainability in Artificial Intelligence-Based Applications: Integrating Analysis and Evaluation (MM4XAI-AE) Models. *Int. J. Intell. Syst.* **2025**, *2025*, 4934696. [CrossRef]