

Mediational Opacity: How AI Reshapes Our Epistemic Environments, and How to Audit It

Alessandro Facchini

SUPSI, Dalle Molle Institute for Artificial Intelligence (IDSIA), Switzerland
alessandro.facchini@supsi.ch

Abstract

AI systems do more than operate within our epistemic environments. They actively reshape them by modifying what users treat as relevant evidence, how they come to know things, and even the standards by which they judge whether a system can be trusted. We argue that this reshaping is largely hidden from view, and we call the resulting condition ‘mediational opacity’. It arises when an AI system whose inner behaviour resists inspection stops being a tool we consciously pick up and use, and instead becomes an environment we inhabit. This type of opacity is different from the two problems most current work addresses, namely the ‘black-box problem’ that explainable AI tries to ‘open up’ and the question of whether individual outputs are fair or accurate, which responsible-AI frameworks track. It is something prior to both, and related precisely to the fact that a deployed system quietly restructures the way its users think and what they count as good judgement. Four reinforcing features (phenomenological silence, dynamic self-reinforcement, societal scale, and contextual promiscuity) make this condition especially hard to detect. Because making AI systems more explainable is not enough to address these features, we propose mediation audits as a complementary approach, focused on the relationship between system, user, and environment, structured along five dimensions, and grounded in a capability approach extended to treat AI as an environmental factor. Finally, we show that agentic AI radicalises the problem since by delegating the ‘thinking’ itself, it produces chains of intermediate epistemic operations that no human can directly observe.

1 Introduction

AI systems do more than process information.¹ They structure the epistemic environments in which humans think, and the practices through which they act. Different authors have recognised this from different perspectives. On the one hand, it has been argued that AI is epistemically distinctive (AI-

varado 2023), that it mediates rather than transmits neutrally (Verbeek 2005, 2011), and that the ‘human-AI hybrid’ it enables can achieve genuine epistemic authority only under specific conditions (Ferrario, Facchini, and Termine 2024). On the other hand, a growing body of work has analysed AI’s opacity, producing both taxonomies (Burrell 2016; Creel 2020; Facchini and Termine 2022) and sociotechnical accounts of opacity in deployment contexts (Mittelstadt et al. 2016; Selbst et al. 2019). These literatures have developed in relative isolation, but taken together, we argue, they point to a shared predicament, which we name *mediational opacity*: AI structures the conditions of human knowing in ways that resist inspection. The condition is, as we will show, structurally different from both the algorithmic opacity that explainable AI addresses and the output-level effects that frameworks for responsible AI deployment track. It concerns indeed how a deployed AI system reshapes the epistemic environment of its users, including what they treat as relevant evidence and the standards by which they evaluate the system.

In this work we make three contributions. First, we develop the concept of mediational opacity, distinguishing it from prior accounts of algorithmic opacity and showing how it arises from the combination of opacity intrinsic to AI’s epistemic operations, and of environmental mediation. Second, we propose mediation audits as a practical framework for assessing how a deployed AI system reshapes its users’ epistemic environment. The audits are structured along five dimensions and grounded in the capability approach to AI ethics. Third, we extend the analysis to agentic AI, showing that delegating the mediation to the system produces a new condition we call *epistemic delegation*, which erodes the conditions for responsible human-AI coupling.

2 AI as Epistemic Mediator

Verbeek’s mediation theory, developed out of Ihde’s post-phenomenology (Ihde 1990) and Latour’s actor-network theory (Latour 1993, 2005), holds that technologies actively co-constitute the relation between human and world, and are therefore never neutral instruments. They mediate perception by shaping how reality is disclosed to us, and action by organising how we engage with that reality, thereby co-shaping what can be perceived and known, as well as the decisions and practices through which we act (Verbeek 2005,

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹In this paper, by AI system we mean a machine learning system, broadly conceived as “any computational artefact capable of learning information from data and adapt its behaviour accordingly” (Termine, Ratti, and Facchini 2026, p.7). This includes predictive systems such as recommenders and diagnostic tools, as well as LLM-based conversational systems.

2011). Aydin, González Woge, and Verbeek (2019) extend this framework to address the fact that contemporary information and communication technologies are increasingly fading from explicit view, embedded in smart environments and ubiquitous computing infrastructures, so that they no longer operate as discrete, object-like tools standing opposite a user, but instead become an intrinsic part of the material environment in which we live. Crucially, such systems are not encountered as discrete instruments that one picks up and sets down, what we refer to as ‘focal tools’, but as continuously running infrastructures woven into the interfaces of everyday devices and the rhythms of daily life, such that users do not so much use the algorithm as inhabit an environment already pre-structured by it. The locus of mediation shifts accordingly from individual tools to the surrounding environment,² a so called ‘mediating milieu’.³

Mediation theory tells us how technologies shape human engagement with the world, but it does not tell us yet what kind of engagement AI specifically shapes. This is where Alvarado’s (2023) analysis is useful. He identifies three dimensions along which a technology can be epistemic: context (designed for inquiry), content (operating on data as evidence), and operations (performing prediction, classification, or inference). A microscope satisfies context and content, but its operations are not epistemic in the relevant sense, magnification being an optical transformation, not a judgement. The microscope makes phenomena visible, whereas the human does the epistemic work. AI, by contrast, satisfies all three dimensions simultaneously and intrinsically. Its operations are epistemic, in that prediction, classification, and inference are knowledge-producing acts whose substantive content emerges from training instead of human specification, and epistemic agency resides in the technology. This is why the ethical concerns generated by algorithmic systems are, as Mittelstadt et al. (2016) observe, epistemological at their core.

Applied to AI, mediation analysis reveals a recurring structure. AI systems amplify the speed, consistency, and scale of decision-making while inhibiting practices on which good judgement depends, including deliberation, sensitivity to context, and skill development (Grote and Berens 2020; Selbst et al. 2019; Vallor 2015; Natali et al. 2025). Prima facie, this amplify/inhibit structure is unremarkable. A speed bump amplifies caution and inhibits velocity in much the same formal sense. What makes AI distinctive as a mediating technology is the conjunction of two features that, e.g., a speed bump or a microscope, each exhibit only in part.

The first is AI’s epistemic character. A speed bump mediates but it does so pragmatically, not epistemically, as it

²Recommender systems exemplify this mode, shaping the user’s epistemic environment by determining what is made visible and what remains hidden.

³The focal-environmental distinction adapts a contrast familiar from postphenomenology and philosophy of technology. At one end, focal tools, in a sense adapted from Borgmann (1984), are discrete objects that users pick up and put down. At the other end, mediating milieux (Aydin, González Woge, and Verbeek 2019) are environments users live in, not tools they use. We treat this contrast as a spectrum along which AI deployments can be situated.

shapes action without producing knowledge claims. A microscope mediates epistemically, but only in a weak sense, since its operations remain human epistemic operations performed through an optical aid. In AI systems, by contrast, the mediating function is itself the performance of epistemic operations (prediction, classification, inference). We call a technology with this property an *epistemic mediator*. Its opacity is therefore qualitatively different from that of other artefacts, as we develop in Section 3.

The second is AI’s environmental character. AI systems increasingly function as mediating milieux rather than as focal tools, with the consequence that their opacity is no longer contained within episodic acts of use but diffuses across the infrastructures that users inhabit. Where the opacity of a focal tool can in principle be ‘bracketed’ (isolated), since one can consult another source or check by hand, the opacity of a mediating milieu resists comparable circumvention. The environment cannot, in fact, be set aside, and any position from which one might audit it is partly constituted by the milieu. We refer to mediation that operates through such a milieu as *environmental mediation*.⁴

In the rest of the paper, we examine the challenge arising from the conjunction of these two threads.

3 Opacity in Epistemic Mediators

Existing taxonomies of algorithmic opacity have advanced our understanding of AI systems largely by focusing on the model’s internal workings and its representational relation to the target phenomenon. Burrell (2016) distinguishes three sources of opacity, the first two social (intentional secrecy, technical illiteracy), and the third computational (intrinsic complexity). Creel (2020) proposes three dimensions of computational transparency (algorithmic, structural, run). Most relevant to the present analysis, Facchini and Termine (2022) distinguish between three forms: ‘access opacity’ (limited epistemic access to the system’s structure and functioning, whether due to secrecy, technical illiteracy, or system complexity), ‘link opacity’ (the inability to establish whether the model genuinely represents the target phenomenon, something closely related to Sullivan’s (2022) ‘link uncertainty’), and ‘semantic opacity’ (the inability to interpret the model’s learned representations in humanly meaningful terms). This latter taxonomy is the most epistemologically fine-grained of the three, since it categorises the epistemic relation between model and user that fails, hence tracking directly the kinds of understanding at stake. Moreover, as it takes into account the relation between the model and the target phenomenon, with link opacity, the taxonomy moves beyond a system-internal analysis. Nevertheless, even this relational step does not reach the question that organises our complementary distinction. Facchini and Termine’s taxonomy concerns the epistemic assessability of the model, whether we can access, verify, and interpret its workings and representations. Ours concerns the epistemic transformation of the user, the inspectability of the model’s me-

⁴See Appendix D of (Facchini 2026), the extended version of this work, for a consolidated overview of the paper’s conceptual framework, including a glossary of key terms.

diation of its users' epistemic environment, including their practices. From this latter perspective, three types emerge:

Accidental opacity. Consider a microscope. Most users cannot explain Abbe diffraction limits. But the mediation structure (what the microscope amplifies and reduces) is fully derivable from the design. A physicist can explain exactly what it shows and hides. The opacity is therefore eliminable in principle, since someone in the epistemic chain has full understanding of the mediation structure.

Epistemically irrelevant opacity. Consider a general-purpose computer. Nearly no one understands what happens between keystroke and screen. But the computer executes a human-specified computational strategy. Every epistemic decision (e.g. which variables to include, or what functional form to assume) was made by the human before the computation began. The system's contribution is speed and accuracy in carrying out that strategy, not judgement about what matters or what to infer. Even when the computation involves information transformation, the transformation follows from instructions that encode human epistemic choices. Opacity here does not threaten knowledge, in the sense that the opaque process is computational execution, not epistemic judgement.⁵ This corresponds roughly to Burrell's 'technical illiteracy', but our point is that such illiteracy is epistemically benign because the opaque process is not an epistemic operation. An AI system, by contrast, does not execute a pre-determined epistemic strategy. Through automated optimisation, the training process discovers which features carry predictive weight, constructs its own representations, and determines decision boundaries, operations that in traditional modelling require domain expertise and theoretical understanding. The epistemic judgements embedded in the learned model emerged from training rather than from human specification, and it is this migration of epistemic agency from the human to the system that makes AI's opacity epistemically constitutive.

Constitutive epistemic opacity. This is the type of opacity that differentiates AI from other technologies. It arises from the convergence of two insights. First, following Alvarado (2023), the opaque process in AI is an epistemic operation, namely a prediction or classification that functions as a knowledge claim. The opacity is constitutive of how the AI system learns its epistemic judgements and how it produces its epistemic outputs. The term 'epistemic' here characterises both the direction of the opacity (opaque to human knowers) and the nature of the opaque process (an epistemic operation). Second, following Facchini and Termine (2022), what is blocked goes beyond cognitive intractability. The model-world relationship resists verification (link opacity), and the learned representations resist interpretation in humanly meaningful terms (semantic opacity). When the opaque process is epistemic and both link and semantic opacity are present, the three blockages reinforce each other

⁵This is not to deny that computer opacity can frustrate knowledge in other senses, for example when a malfunction resists diagnosis. But such cases concern opacity as an obstacle to inquiry about the technology, not opacity as a feature of epistemic operations the technology performs. Our concern is with the latter.

in a way that has no analogue in other computational technologies.

One might object that even if the system's internal epistemic operations are opaque, the effects of its mediation on users could be studied empirically through behavioural observation, outcome analysis, or controlled comparison with unmediated baselines. The objection has force when each component of constitutive epistemic opacity is taken alone, since external validation can compensate for cognitive intractability alone (as in physics simulations), human tracing can compensate for link opacity alone (as in simple regressions), and redesign can compensate for semantic opacity alone (as in manually constructed coding schemes). When the opaque process is epistemic and both link and semantic opacity hold, however, the convergence closes these canonical compensatory routes, because each strategy depends on resources that the other blockages may have made unavailable. Combined with the compounding features developed in Section 4, this convergence can render the mediation practically uninspectable. The predicament is further aggravated by the fact that AI's epistemic outputs typically lack the external checkpoint available to non-epistemic technologies. They concern judgements for which no independent ground truth is accessible, or for which the system's own deployment reshapes the conditions under which verification could occur.⁶

Partial detection of mediation effects remains possible and practically essential, as we argue in Section 6. What empirical observation cannot do is to reconstruct the full characteristic pattern of amplifications, reductions, invitations, and inhibitions that a deployed system produces in a given use context, together with how that pattern evolves over time and varies across population subgroups. We call this pattern the system's *mediation profile*, and the five dimensions of the audit framework address each of its components.

4 Compounding Features

Mediational opacity is produced by the combination of the constitutive epistemic opacity of an epistemic mediator and environmental mediation. The four features developed in this section describe environmental mediation from different analytical angles. Each feature is a dimension along which environmental mediation manifests, and each independently blocks a route by which a focal tool could be isolated and its mediation thereby made visible.

Phenomenological silence. At the level of user awareness, users do not perceive themselves as inhabiting a milieu, and the first condition for critical engagement, being aware that mediation is occurring at all, is systematically undermined. A speed bump announces itself, and an ultrasound is an explicit clinical procedure. AI, however, often mediates without announcing that it is mediat-

⁶An example of the first case is an automated essay score: no independent fact-of-the-matter exists about what the essay 'really' deserves, since judgement varies by evaluator and criteria. An example of the second is a recommender system that reveals certain content: the user's subsequent engagement cannot then be disentangled from the prediction that led to showing it.

ing.⁷ Even where outputs are not presented as definitive, the framing through which they are produced, what is included, excluded, foregrounded, or backgrounded, typically remains invisible. Users cannot critically evaluate a mediation whose structuring they cannot perceive.⁸ Unlike familiar tools, whose mediating effects can be identified by comparing aided and unaided observation, AI's phenomenological silence operates at two distinct levels. In embedded systems such as recommender algorithms or diagnostic tools, users may not realise that mediation is occurring at all. In conversational systems, by contrast, users know they are interacting with an AI, yet the specific character of the mediation, what it amplifies, omits, or frames, remains invisible. The first level is absent in some applications, whereas the second is pervasive across all of them.

Dynamic self-reinforcement. At the level of temporal evolution, the milieu reshapes itself through the behaviour it produces. AI systems undergo continuous modification, including retraining and parameter adjustment, and in doing so the mediation profile shifts without user knowledge. Worse, AI may learn from the behaviour it has already produced. Consider a recommendation system. It shows content, users click it (partly because it was shown), and the system treats the click as evidence of relevance. The mediation thus reshapes itself through its own effects, as the outputs it has produced become the inputs (data and behaviours) on which its subsequent mediation is built. Unlike a focal tool whose mediation profile remains stable between discrete acts of use, the milieu is continuously updated by the very behaviour it conditions. This dynamic has no analogue in classical mediation examples.

Societal scale. At the level of aggregation, AI can mediate millions of epistemic environments simultaneously. A microscope mediates one researcher's perception. AI, instead, produces filter bubbles, homogenisation, and systematic bias that are invisible at the individual level and visible only statistically. Because the mediation operates at societal scale, AI systems raise problems of collective epistemology that individual-level analysis cannot resolve, and governance responses must therefore operate at the population level.

Contextual promiscuity. At the level of domain boundaries, the same milieu pervades multiple domains simultaneously. Classical mediating technologies are domain-specific: a microscope mediates microscopy, an ultrasound mediates obstetric imaging, and the boundaries of their mediation coincide with the boundaries of their use. AI, by contrast, can mediate across multiple epistemic domains through the same underlying architecture, producing a milieu whose reach extends beyond any single task or setting. Consider a large language model deployed simultaneously for clinical documentation, patient communication, and literature review within a hospital. Its mediation of literature search (which studies

it brought to light, which it omits) shapes the clinician's understanding of treatment options, which in turn shapes how clinical notes are written, which feeds back into the training data for future iterations. The mediation profile in one domain cannot be understood independently of its mediation of the others. This means that the mediation cannot be localised, studied, or audited one domain at a time, because effects in one domain interact with effects in others. The empirical alternative to inspecting the mediation, that is studying its effects from the outside, presupposes that the mediation can be isolated within a delimited epistemic task. Contextual promiscuity undermines this presupposition by dissolving the very boundaries on which such isolation depends.

With the compounding features now specified, what mediational opacity is can be stated in full:

***Mediational opacity** is the inability to inspect how a deployed AI system reshapes its users' epistemic environment. It is a property of the system-user-environment relation, arising from the convergence of constitutive epistemic opacity and environmental mediation, and persisting because features such as phenomenological silence, dynamic self-reinforcement, societal scale, and contextual promiscuity block alternative routes to inspection.*

As already announced, this distinction has methodological consequences for how the opacity can and cannot be inspected, which we develop in Section 6.

5 Consequences for the Hybrid Agent

Ferrario, Facchini, and Termine (2024) argued that AI, as it lacks understanding and intellectual virtues, cannot be an epistemic authority or an expert alone, but that a human-AI hybrid can achieve genuine authority or expertise through appropriate 'appropriation'. However our analysis reveals a deep tension within this framework. The hybrid requires the human component to critically evaluate the AI's epistemic contribution, to understand what it amplifies and reduces, what it invites and inhibits. But if AI's mediation is constitutively opaque, phenomenologically silent, and self-reinforcing, the conditions for successful appropriation are very difficult to satisfy. The burden is aggravated by what Freiman (2024) identifies as the 'testimonial gap'. Since AI outputs lack the intentional and normative structure of genuine assertion, the human in the hybrid receives not testimony but information, which must be independently evaluated. However, constitutive epistemic opacity makes such independent evaluation highly problematic.

Mediational opacity has consequences for the hybrid agent at two levels. At the first level, it relocates several established problems, such as responsibility gaps (Matthias 2004), epistemic injustice (Symons and Alvarado 2022), professional autonomy erosion (Grote and Berens 2020), and unaccountable authority (Hauswald 2025; Wolkenstein 2024), from the level of individual outputs to the level of the restructuring of the epistemic environment. In each case, the problem is not that a decision is wrong or unexplainable, but that the evaluative framework within which decisions

⁷Recent evidence in clinical context illustrates this point (Budzyń et al. 2025; Yu et al. 2024).

⁸Phenomenological silence is common across technologies (Ihde 1990; Latour 2005). What distinguishes AI's case is that the recession of an AI system from awareness reshapes what counts as evidence, salience, or relevance, because its mediation is epistemic.

are made has been reshaped without detection. What Fricker (2007) calls ‘hermeneutical injustice’ takes on a new technological form when the interpretative resources needed to challenge a decision are systematically restructured by the decision system. At the second level, mediational opacity produces two consequences specific to the hybrid agent.

The first is **loss of counterfactual imagination**. Appropriation requires the human to conceive how they would reason without the system, in order to critically evaluate its contribution. But prolonged exposure to an opaquely mediating technology gradually eliminates this counterfactual baseline, not through forgetting a procedure but through the restructuring of what competent reasoning feels like. A teacher who has used an automated scoring system for years may no longer be able to articulate what evaluating writing looked like before the system, because the system has redefined which qualities of writing appear salient. This goes beyond deskilling, which concerns the loss of a capacity one knows one had (Vallor 2015; Messeri and Crockett 2024). Loss of counterfactual imagination concerns something different, namely the disappearance of the baseline against which such loss would be recognised. Ferdman (2026) introduces the concept of ‘capacity-hostile environments’ to describe how AI structurally impedes capacity cultivation, but do not identify the specific invisibility of the loss that mediational opacity produces. Mediational opacity produces a loss that is invisible to the person experiencing it. Clinical evidence also supports this claim. Budzyń et al. (2025) found that endoscopists’ unassisted adenoma detection rates declined significantly after routine AI exposure. The practitioners did not lose a procedure they could name. They lost a perceptual capacity they had no reason to believe had changed, precisely because the AI had redefined what adequate detection felt like. The mediation operated at two levels. At the surface, it restructured what the endoscopists perceived, whereas at a deeper level, it restructured their capacity to perceive independently. Both restructurings were invisible to those undergoing them (see Natali et al. (2025) for evidence of analogous dynamics across medical specialities). Without a counterfactual baseline, then, the human component of the hybrid cannot perform the evaluative function that appropriation demands.

The second distinctive consequence is **evaluative circularity**. The criteria by which the human would assess the AI’s epistemic contribution are themselves shaped by the system’s mediation. If a scoring system has gradually restructured what teachers treat as good writing, then teacher agreement with the system’s scores reflects the system’s own epistemic framing, and therefore cannot serve as independent validation.⁹ The hybrid’s internal check, the human evaluating the AI, becomes circular when the human’s evaluative standards have been mediated by the very system under evaluation. Evaluative circularity is thus a structural feature of evaluation conducted from within a mediating milieu that already supplies the criteria by which it would be judged. What it blocks is not evaluation as such, but inde-

pendent validation of those criteria. Prolonged exposure intensifies it without producing it, since the milieu constitutes the evaluative standpoint from the first act of evaluation.

6 Mediation Audits

The Limits of Explainable AI

The compounding features developed in Section 4 raise the question of whether XAI methods (Barredo Arrieta et al. 2020; Guidotti et al. 2019), including post-hoc approaches or more sophisticated ones like mechanistic interpretability (Olah et al. 2020; Bereska and Gavves 2024), causal abstraction (Beckers 2022; Geiger et al. 2024), or concept-based methods (Poeta et al. 2023), can be used to address mediational opacity on their own. We argue that they cannot, because mediation is a property of the system–user–environment relation and not of the system as such, and because the compounding features do not operate at the level of the system’s internal structure. We return to this point once the five dimensions are specified, where the limitations can be stated dimension by dimension. Mediation audits and XAI are therefore complementary, and the framework we introduce below builds on XAI methods without substituting for them.

Mediation Audits: A Complementary Framework

To complement existing algorithmic auditing and impact assessment frameworks, we propose the adoption of *mediation audits*. Conceived as the systematic analyses of what a deployed AI system amplifies, reduces, invites, and inhibits, they extend Verbeek’s mediation analysis methodology (2005, 2011) and the principles of ethical constructive technology assessment (eCTA) developed by Kiran, Oudshoorn, and Verbeek (2015), so as to address the compounding features identified in Section 4. Where Kiran et al. argue for going ‘beyond checklists’ to engage with the normative implications of technological mediation from ‘within’ human–technology relations, we argue that this engagement must be explicitly adapted for epistemic technologies whose mediation is constitutively opaque.

A mediation audit differs from a standard algorithmic audit (Sandvig et al. 2014; Raji et al. 2020) or impact assessment (Metcalf et al. 2021; Selbst 2021) in several respects. Three are worth mentioning. Mediation audits operate on a different unit of analysis than standard algorithmic audits. Where standard audits ask whether the system made a fair or accurate decision, mediation audits ask how the system reshapes what its users perceive, believe, and do. They are also explicitly temporal, tracking how the mediation profile *drifts over time*, including feedback-loop dynamics. Moreover, they address population-level epistemic effects, examining how the system shapes *collective* knowledge across its user base, beyond the individual decisions that standard audits assess.

The shift from system-level to mediation-level analysis carries a broader design implication. If mediational opacity is a property of the relation between the technology, its users, and the epistemic environment, then responsible

⁹A related dynamic has been observed empirically in credit scoring by Millo, Spence, and Xu (2024).

deployment cannot be achieved by designing the technology in isolation. What must be designed, monitored, and governed is the *epistemic ecosystem* in which the technology operates, that is the ensemble of human practices, institutional contexts, interface choices, feedback pathways, and evaluative norms through which the technology's mediation takes concrete form.¹⁰ Before specifying what the audit should examine, we must specify what it should protect within the ecosystem. Without such an anchor, the audit would have no basis for deciding which mediation effects matter, and would become unworkable. To this end, we draw on two complementary accounts. Elgin (2017; 2025) provides the normative content, arguing that epistemic progress depends on autonomy, responsible disagreement, and revisability of understanding, precisely the conditions that the consequences identified in Section 5 systematically disrupt. Ratti and Graves (2025) provide the operational framework through the capability approach, on which AI ethics is operationalised by identifying the human capabilities at stake in a given deployment, determining the conversion factors (personal, social, environmental, and digital) that sustain those capabilities, and assessing how the AI system affects those factors. Our adaptation of this framework extends the standard capability-theoretic usage in one respect. Whereas the capability approach typically treats a technology as a resource whose conversion into a functioning depends on personal, social, environmental, and digital factors, we treat an AI system that functions as a mediating milieu as itself an environmental factor, upstream of the conversion process. With this adaptation in place, a mediation effect raises concern when it undermines the conversion factors on which a relevant epistemic capability depends. In the educational case, the relevant capabilities include independent evaluative judgement, particularly the capacity to recognise unconventional merit, and a pedagogical responsiveness that tracks individual student development. In healthcare, they include independent diagnostic reasoning that draws on narrative and embodied clinical evidence, and the capacity to hold the system's recommendation open to revision.

The five dimensions we now propose operationalise this framework by identifying which amplifications/reductions or invitations/inhibitions affect the conversion factors on which relevant epistemic capabilities depend. For each dimension, we specify its theoretical grounding, the guiding questions it generates, the methods it requires, and the way it goes beyond what existing audit approaches capture. Contextual promiscuity, although it figures as a structural reason in the argument above, does not generate a sixth dimension but poses a cross-cutting challenge for audit scoping that we note at relevant points below.

(1) Epistemic amplification and reduction. This dimension derives directly from Verbeek's perceptual mediation categories. Its guiding question is: *what kinds of evidence*

¹⁰Floridi (2023) argues convergently that AI functions by restructuring the environment in which it operates ('enveloping'); our framework operationalises this insight for the epistemic domain. Auditors are themselves mediated; mediation audits are accordingly reflexive practice rather than view from nowhere. Safeguards are developed in (Facchini 2026, Appendix C).

or reasoning does the system make more salient, and what does it suppress or downrank? Unlike a standard bias audit, which asks e.g. whether the system's outputs are fair across demographic groups, an amplification/reduction analysis asks how the system reshapes the epistemic field within which decisions are made *before* any particular output is generated. Concretely, the analysis proceeds in three steps along the capability-approach structure. It begins by identifying the epistemic capabilities most relevant to the domain. It then determines the conversion factors that have historically sustained each capability, that is the personal, social, environmental, and digital conditions under which practitioners have been able to exercise the capability independently. It concludes by assessing whether the AI system, treated as an environmental factor, preserves, erodes, or restructures those conversion factors.

Consider an automated essay scoring system. A standard audit might test whether the system scores essays equitably across student populations. An amplification/reduction analysis would ask different questions. It would for instance ask whether the system amplifies structural and grammatical features while reducing the salience of originality, and critical depth, or whether it privileges well-represented writing conventions while making unconventional but intellectually ambitious work harder to recognise.

Methodologically, each step in this analysis draws on distinct methods. Capability identification requires domain expertise combined with normative analysis of what the domain treats as competent practice. Conversion factor mapping requires collaboration with sociologists, ethnographers, and domain practitioners who can identify the personal, social, environmental, and digital conditions that historically sustained the capability, and which the AI system may be reshaping. Impact assessment requires comparative analysis of decisions before and after the system's introduction, supplemented by sensitivity analysis ('counterfactual probing') to determine which types of evidence the system treats as probative and which it ignores or discounts.

(2) Epistemic invitation and inhibition. This dimension derives from Verbeek's action-mediation categories. Its guiding question is: *what epistemic activities does the system encourage, facilitate, or make effortless, and what does it discourage, make costly, or foreclose entirely?* Where amplification/reduction concerns what users perceive, invitation/inhibition concerns what users do, how the system shapes epistemic practice.

To illustrate this point, consider the Obermeyer et al. (2019) case. The healthcare algorithm they studied predicted future costs rather than future illness. Because Black patients had historically received less care, they generated lower costs despite being sicker, and were systematically deprioritised. From a standard fairness perspective, the problem is racial bias. From a mediation perspective, this proxy choice has a direct invitation/inhibition structure: the system invited a particular way of thinking about health need (as a function of cost) and inhibited the question of whether cost-generating patterns reflect historical inequities in access. The problem hence lies in how the epistemic field was structured before any output was produced. At the invita-

tion/inhibition level, this structural epistemic redirection is visible from what the system was built to predict, independent of any disparate-impact analysis. It remains visible even where no disparity results. Read through the capability approach, invitation and inhibition shape more than behaviour. They restructure the conversion factors through which epistemic capabilities are exercised. Here, the resulting invitation to allocate resources by predicted cost eroded a conversion factor for the capability of equitable clinical reasoning, namely attention to whether low utilisation reflects unmet need rather than absence of need. A capability-grounded audit registers this erosion as the structural problem, whether or not the disparate impact on Black patients has been measured.

Methods for this dimension include ethnographic observation of how practitioners interact with the system in practice, addressing questions such as: ‘Do they explore alternative hypotheses or accept the system’s suggestions?’ ‘Do they seek additional evidence or treat the system’s output as sufficient?’ Interaction analysis can reveal which epistemic activities the interface facilitates (one-click acceptance, automated workflow integration) and which it makes costly (manual override, justification requirements for deviation). Notice that Selbst et al.’s (2019) warning about ‘abstraction traps’ (that is the tendency to frame algorithmic problems in isolation from their sociotechnical context) applies directly here, since invitation and inhibition emerge from the deployment of the system, not from the algorithm taken alone.

(3) Phenomenological visibility. This dimension addresses the environmental character of AI mediation most directly. The environmental thesis developed in Section 2 holds that AI systems increasingly function as mediating milieux that users inhabit. Phenomenological visibility asks whether users perceive themselves as operating within such a milieu at all. Its guiding question is: *to what extent are users aware that epistemic mediation is occurring, and does the system present its outputs as ‘the answer,’ as a neutral summary, or as one mediated perspective among others?* A bias audit can assess whether outputs are fair across groups without ever asking whether users perceive the outputs as mediated. Dimension 3 asks precisely this question. The answer indicates how far along the focal-environmental spectrum a given deployment has progressed.

This dimension is absent from virtually all existing audit frameworks. Algorithmic audits examine system performance, whereas impact assessments examine downstream effects. Neither asks whether users are phenomenologically aware that mediation is taking place. Nevertheless, awareness is the prerequisite for all critical engagement. A physician who does not realise that the ranked list of diagnoses on her screen has been shaped by a model trained primarily on data from large urban hospitals cannot discount appropriately for a rural patient population, just as a researcher unaware that an AI literature review over-represents recent publications cannot compensate by seeking older foundational work.

Verbeek (2011) proposes what he calls “technology accompaniment”, that is making the mediating role of technology visible and discussable rather than allowing it to re-

cede into the background. Translated into audit methodology, this requires interface analysis (‘does the system signal its own mediation?’), user perception studies (‘do users understand they are interacting with a mediated epistemic environment?’), and comparison with unmediated baselines (‘how would the epistemic task proceed without the system?’).

(4) Temporal stability and feedback dynamics. This dimension addresses the compounding feature of dynamic self-reinforcement. Its guiding question is: *how has the system’s mediation profile changed over time, and are there feedback loops in which the system’s outputs shape the data or user behaviour that in turn reshapes its future mediation?*

This is perhaps the most technically demanding dimension, and the one most clearly absent from existing approaches. As they examine the system at a point in time, standard algorithmic audits are typically snapshot assessments. But as Hatherley (2025) argues, ongoing model updates create update opacity, namely explanations generated for version N of a model may not hold for version $N + 1$. Our framework extends this insight beyond individual model updates to the broader phenomenon of self-reinforcing mediation drift. A recommendation system that shows certain content, observes engagement with that content, and treats the engagement as evidence of preference creates a self-reinforcing loop in which the mediation structure becomes a moving target.

Methods for this dimension include longitudinal monitoring of system outputs across successive model versions and data updates, controlled deployment comparisons that examine how different mediation profiles shaped user behaviour, and drift detection protocols that identify when the relationship between inputs, outputs, and real-world outcomes has shifted significantly. Importantly, temporal audits should extend beyond standard performance metrics (accuracy, fairness over time) to include mediation metrics that operationalise the dimensions as quantifiable trajectories.

(5) Population-level epistemic effects. This dimension addresses the compounding feature of societal scale. Its guiding question is: *what are the aggregate effects of the system’s mediation across its entire user base, and does it systematically advantage some epistemic communities over others?*

Individual-level analysis, including individual explanations, cannot capture population-level epistemic effects. Filter bubbles and epistemic monocultures, along with a systematic homogenisation of inquiry, are emergent properties that arise only when a system mediates millions of epistemic environments simultaneously. Consider a news recommendation algorithm deployed across a national media platform. At the individual level, each user’s feed appears to present ‘relevant news’. The mediating selection (which sources are revealed, which are suppressed, which framings of events are amplified) is invisible to any single user. Only an analysis across the full user base reveals the population-level epistemic effect. Different demographic or geographic groups are systematically shown different content, and the resulting epistemic communities inhabit substantially different informational environments.

Methods for this dimension include distributional analysis ('how do the system's outputs differ across population subgroups in terms of the epistemic resources they make available, alongside differences in accuracy?'), ecological studies ('how does the system's deployment affect the knowledge landscape of a professional community or region?'), and comparative analysis across deployment contexts ('does the same system produce different mediation profiles when deployed in different institutional, cultural, or demographic settings?'). This dimension connects the mediation audit to the broader concerns of social epistemology by asking not just whether the system works fairly for individuals but whether it supports or undermines the conditions for healthy collective knowledge production.

With the five dimensions now specified, we can sharpen the claim introduced at the beginning of this section. Each dimension targets a property of the system–user–environment relation that lies beyond the analytical scope of existing XAI methods. The limitation concerns what XAI is designed to address, not its technical soundness. Indeed, XAI methods deliver essential system-level insights, which mediation audits then situate within the broader epistemic ecology of the system as deployed. Causal XAI can identify which factors drove a particular output, but cannot reveal that the system has systematically restructured what counts as evidence across a domain (dimension 1).¹¹ Concept-based XAI can show which human-meaningful concepts the model uses, but cannot show which concepts the model has rendered invisible by not using them (dimension 2).¹² No XAI method addresses whether users perceive the system as mediating at all (dimension 3), how the mediation profile evolves through feedback loops (dimension 4), or whether epistemic monocultures are forming at the population level (dimension 5).

Minimum Viable Mediation Audit

Because it tracks the environmental character of AI mediation directly, dimension 3 also carries the greatest theoretical weight of the five dimensions. A deployment that scores well on dimension 1 (users can articulate what is amplified and reduced) but poorly on dimension 3 (users do not perceive the mediation as mediation) indicates that the system has progressed along the focal-environmental spectrum in a way that makes external evaluation structurally difficult. Dimension 1 without dimension 3 is diagnostically incomplete, because it describes what the system does without detecting whether users are in a position to recognise that it is doing it. Read through the adapted capability approach, the two dimensions together track whether the AI system, as an environmental factor, is preserving or eroding the conversion

factors sustaining the domain's relevant epistemic capabilities. Dimension 1 identifies whether those factors have been restructured. Dimension 3 identifies whether practitioners can still perceive the restructuring and adjust to it. Temporal drift (dimension 4) cannot be tracked without a baseline amplification/reduction profile (dimension 1). Population-level effects (dimension 5) cannot be identified without knowing what the system makes salient and whether users perceive it as mediated (dimensions 1 and 3). Invitation and inhibition (dimension 2) can be partially inferred from amplification/reduction analysis, though a full assessment requires independent methods.

The minimum protocol¹³ covers dimensions 1 and 3. For dimension 1, identify the system's most heavily weighted features from documentation or XAI tools, then have practitioners rank evidence types by importance, comparing rankings before and after deployment. For dimension 3, three survey items are needed, testing respectively whether the output can still be framed as mediated, whether the conversion factors sustaining prior practice have been preserved, and whether the practitioner can still describe how the task would proceed without the system. A full audit is warranted when practitioners' rankings shift significantly, or when most of them treat the output as an unmediated measure of quality.¹⁴ The absence of these triggers shows only that the most detectable symptoms have not yet appeared.

Dimensions 1 and 4 both presuppose a baseline of practitioner judgement prior to deployment. For systems already in use, such a baseline is typically unavailable, and reconstructing it retrospectively is subject to the loss of counterfactual imagination discussed in Section 5. Partial substitutes may be obtained as follows. On the one hand, comparison with institutions that have not yet adopted a system might provide a synchronic baseline when a diachronic one is absent. On the other hand, cross-institutional variation in mediation profiles could indicate the direction of the shift, even when its starting point cannot be recovered. Clearly, neither is equivalent to a pre-deployment measurement, and in both cases the fact of dealing with approximations should be reported.

Organisations that are unable to carry out even this minimal protocol should explicitly state that no mediation assessment has been performed, and treat this absence as a known epistemic limitation of the deployment. Such a statement makes clear that the organisation is unable to determine whether the system is influencing how users perceive, reason, or act within the relevant domain, and indicates that users and oversight bodies should take this limitation into account when assessing the system and its use.

¹¹ A causal account of why a particular essay received its score does not reveal that, across a domain, the system has gradually shifted teacher attention from originality and critical depth toward structural and grammatical features.

¹² In the Obermeyer et al. (2019) case, concept-based XAI could reveal 'prior costs' as a concept the algorithm uses, but not the fact that this reliance invited thinking about health need as cost-dependent and inhibited thinking about it as independent of utilisation history.

¹³ The protocol is provided in (Facchini 2026, Appendix B).

¹⁴ Facchini (2026, Appendix A) outlines the application of the five dimensions to an automated essay scoring system, specifying, for each dimension, the capability at stake, the methods required, and the indicators that would raise concern. Other domains such as healthcare, criminal justice, scientific research, and news recommendation would each require domain-appropriate instruments but follow the same five-dimension structure.

Positioning and Criteria for Adequacy

Mediation audits supplement existing frameworks without replacing them, since they assess a different object. Standard audits, in fact, examine what a system produces, whereas mediation audits ask how what it produces reshapes practitioners' standards for evidence and inference. The difference matters, because a system can pass standard accuracy and fairness checks while having nonetheless restructured what practitioners treat as evidence, and no inspection of outputs will detect this.

What counts as an adequate mediation audit, given that mediational opacity cannot be eliminated? We propose four success criteria. The first criterion is *contestability*. The audit should produce findings that enable affected parties to challenge not only specific outputs but the epistemic framing the system imposes, what it treats as evidence, what it renders invisible, and what practices it invites. The second criterion is *drift detection*. The audit should establish baseline mediation profiles against which later shifts can be identified while they are still reversible. The third criterion is *visibility thresholds*. The audit should determine whether users' awareness of mediation falls below the level required for the critical engagement that responsible coupling demands, and flag deployments where it does. The fourth criterion is *capability preservation*. The audit should assess whether the conversion factors sustaining the epistemic capabilities relevant to the domain remain intact under the system's mediation, and flag deployments in which those factors have been eroded below the threshold required for independent exercise of the capability. Together, these criteria constitute the minimum conditions under which mediation audits can meaningfully support epistemic responsibility in the absence of full transparency.¹⁵

7 Agentic AI and Epistemic Delegation

The analysis so far has focused on AI systems where the human remains in the epistemic loop, even if their perception is shaped. The rapid emergence of agentic AI systems (autonomous agents that plan, use tools, and execute multi-step tasks without continuous human oversight (Chan et al. 2024; Gabriel et al. 2024)) introduces a structural transformation that our framework is positioned to illuminate. In standard AI, the human experiences the world as filtered through the system's amplifications and reductions. In agentic AI, the human does not experience a mediated perception at all for large portions of the task. The agent autonomously searches, evaluates sources, decides what is relevant, formulates sub-goals, selects tools, acts on its decisions, observes results, and adjusts.¹⁶ The human receives the output of an entire

¹⁵Facchini (2026, Appendix C) situates mediation audits with respect to existing algorithmic auditing work, develops specific integrations with eCTA (Kiran, Oudshoorn, and Verbeek 2015), algorithmic impact assessments (Metcalf et al. 2021; Selbst 2021), and Value Sensitive Design (Friedman and Hendry 2019), and addresses institutional conditions for implementation and the reflexive challenge posed by auditors' own mediated positions.

¹⁶We describe the agent's operations in functional terms, without imputing subject-like intentionality. These are computational

chain of epistemic operations from which they were absent. This is not a departure from mediation. All technologies mediate, and the human still stands in a mediated relation to the world through the agent's output.¹⁷ But the *structure* of that mediation has changed categorically. In standard AI, the human experiences a single mediating layer whose amplifications and reductions, while opaque, are at least undergone. In agentic AI, the human receives the endpoint of a chain of mediations none of which were experienced by anyone. The intermediate epistemic operations (e.g., searching, evaluating, selecting, interpreting) were performed autonomously, each shaping the epistemic environment for the next. We call this phenomenon *epistemic delegation*, a condition in which the mediating process itself is delegated. Mediation continues, though the human can no longer inspect, evaluate, or even perceive the intermediate mediations that shaped the output they receive.

Situating agentic AI on the focal-environmental spectrum introduced in Section 2 clarifies the transformation at issue. Standard AI systems already extend beyond the focal end of this spectrum. Even a chatbot, while formally a discrete tool, operates environmentally through its integration into workflows and interfaces, and through patterns of habitual use that induce phenomenological silence. Agentic AI intensifies this trajectory. As intermediate epistemic operations are delegated to the agent, users no longer inhabit the epistemic milieu throughout the task. This withdrawal from the chain of epistemic operations marks the extreme point of a continuum along which AI systems have progressively redistributed epistemic activity.

One might object that reasoning traces (intermediate reasoning steps) restore inspectability, but the trace is itself a mediated summary, presents only the path taken and not the paths rejected, and would require full epistemic evaluation of every intermediate step to serve its purpose, which negates the efficiency that motivated delegation. The consequence for Ferrario, Facchini, and Termine's (2024) hybrid agent is severe. The hybrid requires in fact the human to critically evaluate the AI's epistemic contribution. But when the human has not witnessed the chain of intermediate mediations that produced the output, the conditions for appropriation are structurally undermined for the delegated portions of the chain. Moreover, each step's mediation shapes the epistemic environment for the next, producing a chain-level opacity that no explanation of the final output can recover.

Phenomenological silence in a sense reaches its limit here. The user is not merely unaware that mediation is occurring; they are not present during the epistemic operations at all. The agent's central appeal, freeing the user from epistemic labour, is precisely what makes responsible epistemic coupling hardest to sustain. As more epistemic work is delegated to the agent, the human correspondingly exercises

processes that perform the epistemic functions that in non-agentic AI are performed by the human user.

¹⁷We use 'delegation' instead of 'replacement' deliberately. Agentic AI does not eliminate mediation but delegates the mediating process itself, the chain of epistemic operations through which the final output is produced.

those evaluative capacities less often. Over time, this erosion of practice makes it increasingly difficult for the human to assess the agent's outputs (cf. Messeri and Crockett (2024) on 'illusions of understanding'), generating a self-reinforcing dynamic in which the human's contribution to the hybrid progressively weakens.

Agentic systems hence require mediation audits that go beyond the five dimensions, addressing chain-level opacity, delegation boundaries, and specific deskilling trajectories. Epistemic delegation denotes a structural transformation that alignment and control framings cannot capture. This paper specifies that transformation and the audit dimensions it entails, whereas governance instruments remain beyond its current scope.

8 Related Work

Some relevant antecedents were addressed throughout. Friedrich, Mason, and Malone (2022) offer the closest antecedent, applying postphenomenology to clinical AI. We extend their analysis by developing the constitutive epistemic opacity concept, generalising beyond the clinical context, and proposing the mediation audit framework. Susser (2019) focuses on autonomy over epistemology. Our framework reinterprets his 'invisible influence', and Aydin, González Woge, and Verbeek's 2019 account of AI's tendency to fade into an ambient 'technological environmentality' as instances of phenomenological silence. That mediation is environmental is a claim with a long tradition in media studies and media ecology (McLuhan 1964). In a similar move, Pasquale (2015) located opacity in the 'socio-institutional arrangement' around the model, rather than in the model as such. Hatherley (2025) has identified a problem related to dynamic self-reinforcement, 'update opacity', whereby ongoing model updates create a moving target that undermines even those forms of understanding that XAI methods have achieved. Ratti and Graves (2025) propose a capability approach to AI ethics that identifies the human capabilities at stake in a given deployment and the conversion factors that sustain them, an approach then operationalised by Graves and Ratti (2025) for internal auditing of the design and development process. Our mediation audit framework adopts this approach as its operational anchor and extends it both conceptually, by repositioning the AI from resource to environmental factor in light of its character as a mediating milieu, and methodologically, by specifying through the five dimensions how the AI so positioned affects conversion factors at the level of perception, practice, and epistemic environment. Empirical work on human-AI decision-making supplies evidence for specific mediational dynamics without framing them in mediation-theoretic terms (Skitka, Mosier, and Burdick 1999; Buccinca, Malaya, and Gajos 2021; Kaur et al. 2020; Rezazade Mehrizi et al. 2023; Dratsch et al. 2023), and supports the claim that mediational opacity operates at a level not addressed by providing explanations or cultivating critical attitudes. On agentic AI, Chan et al. (2024) and Gabriel et al. (2024) identify accountability gaps but frame them in terms of alignment and safety, instead of epistemic delegation.

9 Conclusion

Mediational opacity undermines the conditions for responsible human-AI coupling. We cannot use our epistemic capacities to evaluate the system shaping them, because no standpoint from which to evaluate it stands independent of the milieu. Given that the opaque process is an epistemic operation, and link and semantic opacity prevent compensatory understanding, this is a different kind of opacity from the one found in microscopes or computers. The compounding features describe this environmental condition in multiple ways. The practical implications concern the creation of responsibility gaps and epistemic injustice, which XAI alone cannot resolve. We have therefore proposed mediation audits structured along five dimensions.

Agentic AI further transforms the predicament. When the mediating process is delegated, the human encounters only the endpoint of an epistemic chain whose intermediate operations went unwitnessed, rendering the conditions of appropriation structurally unmeetable for the delegated portions. This is a problem of epistemic delegation, extending beyond alignment and control to the conditions under which epistemic understanding and responsibility can be meaningfully exercised. Addressing it requires dimensions beyond the five specified here. Nor do mediation audits eliminate mediational opacity in the non-agentic case. Epistemic opacity persists as long as link and semantic opacity remain, and environmental mediation endures wherever AI functions as an epistemic milieu rather than a focal tool. Mediation audits reframe the question from whether individual outputs can be explained to whether the mediating process can be made visible and contestable, thereby enabling the sustained critical engagement responsible human-AI coupling requires.

Several questions remain open. One concerns the possibility that users may be unable to perceive the mediation they are subject to, thereby structurally undermining the conditions for meaningful consent, and raising challenges for frameworks such as the EU AI Act that rely on consent and transparency as legitimating mechanisms. Another issue is the potential homogenisation produced when millions of epistemic environments are mediated through shared architectures, which may reduce the epistemic diversity on which collective knowledge depends (Fazelpour and Steel 2022). A further priority is empirical validation through replicable protocols that preserve qualitative depth while enabling cross-site comparability, particularly for dimensions 1 and 3. The framework also generates testable predictions amenable to experimental design, such as the hypothesis that XAI explanations do not significantly reduce the shift in evidence-type reliance a system produces. Mediation metrics require systematic development, such as the ratio of unmodified acceptances to overrides, or the convergence of top-ranked outcomes across retraining cycles, validated against independent evidence of capability erosion. Finally, the governance of agentic AI under epistemic delegation remains underdeveloped.

As AI systems become more autonomous and more deeply embedded in the epistemic infrastructure of society, the need for tools that make not just outputs but mediation visible and contestable will only grow.

Acknowledgments

The work of the author was partly conducted within the framework of the EUonAIR Centre of Excellence in Responsible AI and Education. These activities were partially supported by a grant from Movetia, funded by the Swiss Confederation. The author is grateful to Andrea Ferrario, Emanuele Ratti and Elisa Rubegni, as well as all three AIES reviewers, for their helpful comments and insightful feedbacks on earlier versions of this work.

References

- Alvarado, R. 2023. AI as an epistemic technology. *Science and Engineering Ethics*, 29(5): 32.
- Aydin, C.; González Woge, M.; and Verbeek, P.-P. 2019. Technological environmentality: Conceptualizing technology as a mediating milieu. *Philosophy & Technology*, 32(2): 321–338.
- Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Benetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; Chatila, R.; and Herrera, F. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58: 82–115.
- Beckers, S. 2022. Causal explanations and XAI. In *Conference on causal learning and reasoning*, 90–109. PMLR.
- Bereska, L.; and Gavves, E. 2024. Mechanistic Interpretability for AI Safety: A Review. *Transactions on Machine Learning Research*.
- Borgmann, A. 1984. *Technology and the Character of Contemporary Life: A Philosophical Inquiry*. Chicago: University of Chicago Press.
- Buccinca, Z.; Malaya, M. B.; and Gajos, K. Z. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–21.
- Budzyń, K.; Romańczyk, M.; Kitala, D.; Kołodziej, P.; Bugajski, M.; Adami, H. O.; Blom, J.; Buszkiewicz, M.; Halvorsen, N.; Hassan, C.; Romańczyk, T.; Holme, Ø.; Jarus, K.; Fielding, S.; Kunar, M.; Pellise, M.; Pilonis, N.; Kamiński, M. F.; Kalager, M.; Bretthauer, M.; and Mori, Y. 2025. Endoscopist deskillling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *The Lancet Gastroenterology & Hepatology*, 10(10): 896–903.
- Burrell, J. 2016. How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big data & society*, 3(1): 2053951715622512.
- Chan, A.; Ezell, C.; Kaufmann, M.; Wei, K.; Hammond, L.; Bradley, H.; Bluemke, E.; Rajkumar, N.; Krueger, D.; Kolt, N.; et al. 2024. Visibility into AI agents. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*, 958–973.
- Creel, K. A. 2020. Transparency in complex computational systems. *Philosophy of Science*, 87(4): 568–589.
- Dratsch, T.; Chen, X.; Rezazade Mehrizi, M.; Kloeckner, R.; Mähringer-Kunz, A.; Püsken, M.; Baeßler, B.; Sauer, S.; Maintz, D.; and Pinto dos Santos, D. 2023. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology*, 307(4): e222176.
- Elgin, C. Z. 2017. *True enough*. Cambridge, MA: MIT press.
- Elgin, C. Z. 2025. *Epistemic Ecology*. Cambridge, MA: MIT Press.
- Facchini, A. 2026. Mediational Opacity: How AI Reshapes Our Epistemic Environments, and How to Audit It. *Available at SSRN* 6808278.
- Facchini, A.; and Termine, A. 2022. Towards a Taxonomy for the Opacity of AI Systems. In Müller, V. C., ed., *Philosophy and Theory of Artificial Intelligence 2021*, volume 63 of *Studies in Applied Philosophy, Epistemology and Rational Ethics*, 73–89. Cham: Springer.
- Fazelpour, S.; and Steel, D. 2022. Diversity, trust, and conformity: A simulation study. *Philosophy of Science*, 89(2): 209–231.
- Ferdman, A. 2026. AI deskilling is a structural problem. *AI & SOCIETY*, 41(4): 3001–3013.
- Ferrario, A.; Facchini, A.; and Termine, A. 2024. Experts or Authorities? The Strange Case of the Presumed Epistemic Superiority of Artificial Intelligence Systems: A. Ferrario et al. *Minds and Machines*, 34(3): 30.
- Floridi, L. 2023. *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*. Oxford University Press.
- Freiman, O. 2024. AI-Testimony, conversational AIs and our anthropocentric theory of Testimony. *Social Epistemology*, 38(4): 476–490.
- Fricker, M. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.
- Friedman, B.; and Hendry, D. G. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press.
- Friedrich, A. B.; Mason, J.; and Malone, J. R. 2022. Re-thinking explainability: toward a postphenomenology of black-box artificial intelligence in medicine. *Ethics and Information Technology*, 24(1): 8.
- Gabriel, I.; Manzini, A.; Keeling, G.; Hendricks, L. A.; Rieser, V.; Iqbal, H.; Tomašev, N.; Ktena, I.; Kenton, Z.; Rodriguez, M.; et al. 2024. The Ethics of Advanced AI Assistants. arXiv:2404.16244.
- Geiger, A.; Wu, Z.; Potts, C.; Icard, T.; and Goodman, N. 2024. Finding Alignments Between Interpretable Causal Variables and Distributed Neural Representations. In *Conference on Causal Learning and Reasoning*, 160–187. PMLR.
- Graves, M.; and Ratti, E. 2025. A capability approach to ethical development and internal auditing of AI technology. *Journal of Responsible Technology*, 22: 100121.

- Grote, T.; and Berens, P. 2020. On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics*, 46(3): 205–211.
- Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2019. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51(5): 1–42.
- Hatherley, J. 2025. A moving target in AI-assisted decision-making: dataset shift, model updating, and the problem of update opacity: J. Hatherley. *Ethics and Information Technology*, 27(2): 20.
- Hauswald, R. 2025. Artificial epistemic authorities. *Social Epistemology*, 39(6): 716–725.
- Ihde, D. 1990. *Technology and the lifeworld: From garden to earth*. Indiana University Press.
- Kaur, H.; Nori, H.; Jenkins, S.; Caruana, R.; Wallach, H.; and Wortman Vaughan, J. 2020. Interpreting Interpretability: Understanding Data Scientists’ Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.
- Kiran, A. H.; Oudshoorn, N.; and Verbeek, P.-P. 2015. Beyond checklists: toward an ethical-constructive technology assessment. *Journal of responsible innovation*, 2(1): 5–19.
- Latour, B. 1993. *We Have Never Been Modern*. Cambridge, MA: Harvard University Press. Translated by Catherine Porter.
- Latour, B. 2005. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.
- Matthias, A. 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3): 175–183.
- McLuhan, M. 1964. *Understanding media: The extensions of man*. McGraw-Hill.
- Messeri, L.; and Crockett, M. J. 2024. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002): 49–58.
- Metcalfe, J.; Moss, E.; Watkins, E. A.; Singh, R.; and Elish, M. C. 2021. Algorithmic impact assessments and accountability: The co-construction of impacts. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 735–746.
- Millo, Y.; Spence, C.; and Xu, R. 2024. Algorithmic self-referentiality: How machine learning pushes calculative practices to assess themselves. *Accounting, Organizations and Society*, 113: 101567.
- Mittelstadt, B. D.; Allo, P.; Taddeo, M.; Wachter, S.; and Floridi, L. 2016. The ethics of algorithms: Mapping the debate. *Big data & society*, 3(2): 2053951716679679.
- Natali, C.; Marconi, L.; Dias Duran, L. D.; and Cabitza, F. 2025. AI-induced deskilling in medicine: a mixed-method review and research agenda for healthcare and beyond. *Artificial Intelligence Review*, 58(11): 356.
- Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.
- Olah, C.; Cammarata, N.; Schubert, L.; Goh, G.; Petrov, M.; and Carter, S. 2020. Zoom in: An introduction to circuits. *Distill*, 5(3): e00024–001.
- Pasquale, F. 2015. *The black box society: The secret algorithms that control money and information*. Harvard University Press.
- Poeta, E.; Ciravegna, G.; Pastor, E.; Cerquitelli, T.; and Baralis, E. 2023. Concept-based explainable artificial intelligence: A survey. *ACM Computing Surveys*.
- Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 33–44.
- Ratti, E.; and Graves, M. 2025. A Capability Approach to AI Ethics. *American Philosophical Quarterly*, 62(1): 1–16.
- Rezazade Mehrizi, M. H.; Mol, F.; Peter, M.; Ranschaert, E.; Dos Santos, D. P.; Shahidi, R.; Fatehi, M.; and Dratsch, T. 2023. The impact of AI suggestions on radiologists’ decisions: a pilot study of explainability and attitudinal priming interventions in mammography examination. *Scientific reports*, 13(1): 9230.
- Sandvig, C.; Hamilton, K.; Karahalios, K.; and Langbort, C. 2014. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. In *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*. Preconference at the 64th Annual Meeting of the International Communication Association.
- Selbst, A. D. 2021. An institutional view of algorithmic impact assessments. *Harv. JL & Tech.*, 35: 117.
- Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, 59–68.
- Skitka, L. J.; Mosier, K. L.; and Burdick, M. 1999. Does Automation Bias Decision-Making? *International Journal of Human-Computer Studies*, 51(5): 991–1006.
- Sullivan, E. 2022. Understanding from Machine Learning Models. *The British Journal for the Philosophy of Science*, 73(1): 109–133.
- Susser, D. 2019. Invisible influence: Artificial intelligence and the ethics of adaptive choice architectures. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 403–408.
- Symons, J.; and Alvarado, R. 2022. Epistemic injustice and data science technologies. *Synthese*, 200(2): 87.
- Termine, A.; Ratti, E.; and Facchini, A. 2026. Machine learning and theory-ladenness: a phenomenological account. *Synthese*, 207(3): 94.
- Vallor, S. 2015. Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character. *Philosophy & Technology*, 28(1): 107–124.
- Verbeek, P.-P. 2005. *What Things Do: Philosophical Reflections on Technology, Agency, and Design*. Penn State University Press.

Verbeek, P.-P. 2011. *Moralizing Technology: Understanding and Designing the Morality of Things*. University of Chicago Press.

Wolkenstein, A. 2024. Healthy mistrust: Medical black box algorithms, epistemic authority, and preemptionism. *Cambridge Quarterly of Healthcare Ethics*, 33(3): 370–379.

Yu, F.; Moehring, A.; Banerjee, O.; Salz, T.; Agarwal, N.; and Rajpurkar, P. 2024. Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nature Medicine*, 30(3): 837–849.