EthAIca. 2025; 4:417
doi: 10.56294/ai2025417

REVIEW

Epistemic Injustice in Generative AI: A Pipeline Taxonomy, Empirical Hypotheses, and Stage-Matched Governance

Injusticia epistémica en la IA generativa: una taxonomía de la cadena de valor, hipótesis empíricas y gobernanza alineada por etapas

Joffrey Baeyaert[1]

[1]Independent Researcher. Lisbon, Portugal.

Cite as: Baeyaert J. Epistemic Injustice in Generative AI: A Pipeline Taxonomy, Empirical Hypotheses, and Stage-Matched Governance. EthAIca. 2025; 4:417. https://doi.org/10.56294/ai2025417

Submitted: 23-02-2025 **Revised:** 31-05-2025 **Accepted:** 25-08-2025 **Published:** 26-08-2025

Editor: PhD. Rubén González Vallejo

Corresponding Author: Joffrey Baeyaert

ABSTRACT

Introduction: generative AI systems increasingly influence whose knowledge is represented, how meaning is framed, and who benefits from information. However, these systems frequently perpetuate epistemic injustices—structural harms that compromise the credibility, intelligibility, and visibility of marginalized communities.

Objective: this study aims to systematically analyze how epistemic injustices emerge across the generative AI pipeline and to propose a framework for diagnosing, testing, and mitigating these harms through targeted design and governance strategies.

Method: a mutually exclusive and collectively exhaustive (MECE) taxonomy is developed to map testimonial, hermeneutical, and distributive injustices onto four development stages: data collection, model training, inference, and dissemination. Building on this framework, four theory-driven hypotheses (H1-H4) are formulated to connect design decisions to measurable epistemic harms. Two hypotheses—concerning role-calibrated explanations (H3) and opacity-induced deference (H4)—are empirically tested through a PRISMA-style meta-synthesis of 21 behavioral studies.

Results: findings reveal that AI opacity significantly increases deference to system outputs (effect size d ≈ 0,46-0,58), reinforcing authority biases. In contrast, explanations aligned with stakeholder roles enhance perceived trustworthiness and fairness (d ≈ 0,40-0,84). These effects demonstrate the material impact of design choices on epistemic outcomes.

Conclusions: epistemic justice should not be treated as a post hoc ethical concern but as a designable, auditable property of AI systems. We propose stage-specific governance interventions—such as participatory data audits, semantic drift monitoring, and role-sensitive explanation regimes—to embed justice across the pipeline. This framework supports the development of more accountable, inclusive generative AI.

Keywords: Epistemic Injustice; Generative AI Governance; Pipeline Taxonomy.

RESUMEN

Introducción: los sistemas de IA generativa influyen cada vez más en qué conocimiento se representa, cómo se enmarca el significado y quién se beneficia de la información. Sin embargo, estos sistemas con frecuencia perpetúan injusticias epistémicas: daños estructurales que comprometen la credibilidad, la inteligibilidad y la visibilidad de las comunidades marginadas.

Objetivo: este estudio busca analizar sistemáticamente cómo surgen las injusticias epistémicas en el proceso de IA generativa y proponer un marco para diagnosticar, evaluar y mitigar estos daños mediante estrategias de diseño y gobernanza específicas.

Método: se desarrolla una taxonomía mutuamente excluyente y colectivamente exhaustiva (MECE) para mapear las injusticias testimoniales, hermenéuticas y distributivas en cuatro etapas de desarrollo: recopilación de datos, entrenamiento de modelos, inferencia y difusión. A partir de este marco, se formulan cuatro hipótesis teóricas (H1-H4) para vincular las decisiones de diseño con daños epistémicos medibles. Se prueban empíricamente dos hipótesis —relativas a las explicaciones calibradas por roles (H3) y a la deferencia inducida por la opacidad (H4)— a través de una metasíntesis estilo PRISMA de 21 estudios conductuales.
Resultados: los hallazgos revelan que la opacidad de la IA aumenta significativamente la deferencia hacia los resultados del sistema (tamaño del efecto d ≈ 0,46-0,58), lo que refuerza los sesgos de autoridad. Por el contrario, las explicaciones alineadas con los roles de las partes interesadas mejoran la confiabilidad y la equidad percibidas (d ≈ 0,40-0,84). Estos efectos demuestran el impacto sustancial de las decisiones de diseño en los resultados epistémicos.
Conclusiones: la justicia epistémica no debe considerarse una preocupación ética a posteriori, sino una propiedad de los sistemas de IA que se puede diseñar y auditar. Proponemos intervenciones de gobernanza específicas para cada etapa —como auditorías participativas de datos, monitoreo de la deriva semántica y regímenes de explicación sensibles a los roles— para integrar la justicia en todo el proceso de desarrollo. Este marco apoya el desarrollo de una IA generativa más responsable e inclusiva.

Palabras clave: Injusticia Epistémica; Gobernanza de la IA Generativa; Taxonomía de Procesos de Desarrollo.

INTRODUCTION

How, where, and why do generative artificial intelligence (GenAI) systems reproduce epistemic injustice across their life-cycle? This question has become increasingly urgent. GenAI models now produce news articles, illustrate textbooks, and draft legal documents. In less than half a decade, they have shifted from being dismissed as “stochastic parrots” to serving as pivotal epistemic agents that shape public knowledge.(1,2,3) Their capacity to generate text, audio, and images at near-zero marginal cost grants them significant agenda-setting power. Each synthetic output can influence judgments about truth, expertise, and legitimacy.(4,5,6)

The epistemic consequences of these systems extend beyond accuracy. GenAI mediates who is heard, how meaning is structured, and what becomes visible in public discourse.(7,8,9) Yet much existing research continues to frame bias as a localized technical error, rather than as a structural effect of design.(5,10,11) Despite extensive audits of training corpora and attempts to filter outputs, no framework currently links epistemic harms to specific design choices across the AI pipeline.(4,6,7)

This study addresses that gap. It conceptualizes GenAI as an epistemic assembly line in which decisions about data, training, inference, and dissemination allocate credibility, shape interpretive resources, and distribute informational benefits and harms.(6,12) By mapping these pipeline stages onto testimonial, hermeneutical, and distributive injustices, the study introduces a mutually exclusive and collectively exhaustive taxonomy.(4,6) It then advances four theory-driven hypotheses and stage-specific governance interventions—from participatory corpus design to role-calibrated explanations.(7,13,14) These recommendations resonate with emerging regulatory frameworks such as the EU AI Act and underscore epistemic justice not as a secondary concern, but as a structural principle of responsible AI design.(4,6,13)

Conceptual Foundations

Epistemic injustice refers to systematic harm inflicted on individuals in their capacity as knowers (table 1). First articulated by Miranda Fricker, it captures how credibility, intelligibility, and access to knowledge are unequally distributed.(12) Three main forms are widely recognized:

- Testimonial injustice occurs when a speaker’s credibility is unfairly deflated due to prejudice.
- Hermeneutical injustice arises when collective interpretive resources fail to make marginalized experiences intelligible.
- Distributive injustice refers to the uneven allocation of epistemic goods—such as information, visibility, or authoritativeness—across social groups.(4)

These are not incidental errors but structural effects of power, determining who is heard, what is understood, and whose knowledge circulates. In the context of GenAI, these dynamics take on heightened importance. Generative models filter, rephrase, and generate symbolic content, often magnifying the hierarchies embedded in their training data, architectures, and deployment environments.(4,7)

Recent scholarship extends epistemic injustice theory to algorithmic epistemic injustice, reframing GenAI not merely as a technical artefact but as an epistemic agent that participates in knowledge formation and legitimation.(4,15) When models discount inputs from marginalized dialects, they enact testimonial injustice.(5,7)

When they fail to register minority concepts or erase cultural terms, they perpetuate hermeneutical injustice.[6] When their benefits accrue disproportionately to dominant groups, they generate distributive injustice.[11]

These harms emerge across all pipeline stages:

- Data collection: curatorial choices exclude marginalized voices, producing testimonial deficits.[6,16]
- Model training: inductive biases distort minority concepts, codifying hermeneutical gaps.[17,18]
- Inference: hallucinations and stereotypes disproportionately discredit vulnerable groups.[4,19]
- Dissemination: platform dynamics amplify dominant narratives, silencing minority perspectives.[11]

Opacity compounds these harms. GenAI systems often resist interpretation even by their creators, fostering opacity-induced deference: users trust outputs they cannot verify. In high-stakes contexts like medicine, law, and education, this erodes human judgment while reinforcing inequities under the guise of neutrality.[5,20] Reliability alone does not resolve this issue, as epistemic authority remains unequally distributed; marginalized groups are least able to contest or contextualize AI outputs.[6,21,22,23]

Addressing these problems requires more than technical fairness metrics or transparency checklists. A systematic, life-cycle-oriented framework is needed to connect design decisions to epistemic outcomes. This study develops such a framework by mapping each pipeline stage to distinct forms of epistemic injustice and by proposing a taxonomy of harms.[24,25,26] The goal is to transform philosophical concerns into empirically testable claims and governance-relevant strategies. By demonstrating how credibility penalties, epistemic erasures, and opacity-induced deference stem from concrete design choices, the framework positions epistemic justice as a designable and governable feature of generative AI systems.[27,28,29]

Table 1. Epistemic Injustice in Generative AI by Pipeline Stage

Pipeline Stage	Key Epistemic Injustice	Description & Example
Data Collection	Testimonial injustice (silencing by omission)	Under-representation of certain groups means the model “learns” an incomplete worldview. Example: A model trained predominantly on Western English struggles with African American Vernacular English or Indigenous languages, effectively silencing those dialects.[6,7]
Model Training	Hermeneutical injustice (skewed concepts)	Biased learning erases or distorts minority concepts. Example: Fine-tuning erases culturally specific terms, privileging dominant interpretations.[6]
Inference/Output	Credibility deficits & misinformation	Hallucinations and stereotypes disproportionately misrepresent marginalized groups. Example: A model confidently relays a false historical narrative favoring a dominant group.[4,19]
Dissemination	Distributive injustice (unequal reach)	Platforms amplify dominant voices while minority knowledge remains hidden. Example: Multilingual outputs exist but English content is preferentially promoted.[11,15]

METHOD

We used a mixed-methods, exploratory design integrating (i) conceptual theory-building, (ii) illustrative case construction, and (iii) a secondary synthesis of empirical evidence. These components are analytically distinct yet mutually reinforcing: the conceptual analysis establishes a stage-sensitive taxonomy and derives four falsifiable hypotheses (H1–H4); the cases operationalize mechanisms in realistic deployments; and the evidence synthesis aggregates independent behavioral studies from HCI, education, and organizational science to evaluate H3 and H4. Table 2 summarizes the design.

This study combines conceptual inquiry with empirical evidence to trace how different forms of epistemic injustice arise and can be mitigated across the generative-AI life-cycle. We integrate (i) a taxonomy-building literature analysis, (ii) two detailed illustrative cases, and (iii) a structured synthesis of the best available behavioural evidence.

Table 2. Design overview

Methodological Component	Implementation Details	Purpose in Study
Conceptual Analysis (i)	Conducted an iterative literature review across epistemology, AI ethics, and sociotechnical HCI. Mapped the three core types of epistemic injustice, testimonial, hermeneutical, distributive, onto the four canonical stages of the generative AI pipeline: (1) data collection, (2) model training, (3) inference, and (4) dissemination. This produced a mutually exclusive and collectively exhaustive (MECE) taxonomy and informed the construction of four theory-driven hypotheses (H1–H4).[6,11,12]	Provides a logically complete and stage-sensitive theoretical scaffold. Translates abstract philosophical concepts into structured categories that predict observable model behaviors and systemic harms. Forms the basis for hypothesis formulation and later validation.
Illustrative Cases (ii)	Developed two fictional but evidence-grounded narrative scenarios to concretely demonstrate how epistemic injustices manifest in generative AI systems. Case 1: Diagnostic chatbot in healthcare, used to illustrate opacity-induced deference (H4). Case 2: Multilingual news generator, used to show testimonial silencing and dissemination asymmetries (H1, H2). Designs were guided by prior failure modes documented in bias audits.[6,19,22] Each case aligns harms to pipeline stages and suggests corresponding governance levers.	Clarifies the practical stakes of H1–H4 for interdisciplinary readers. Makes abstract harms visible and actionable through high-fidelity system narratives. Supports stakeholder comprehension, without presenting fictional accounts as empirical evidence. Labeled explicitly as illustrative.
Empirical Evidence Synthesis (iii)	Executed a PRISMA-style systematic review across four databases (see Appendix): Scopus, Web of Science, arXiv, and ACM DL (cut-off date: 31 May 2025). Out of 42 initially retrieved records, 21 met pre-registered inclusion criteria (e.g., $N \geq 40$, clear method, dispersion statistics reported). Quantitative studies were converted to Cohen's *d* effect sizes following a formal conversion protocol detailed in the Appendix. Inputs included means with SD, proportions, x^2, η^2, or ANOVA values. When dispersion statistics were missing, studies were excluded from meta-analysis but retained for qualitative synthesis.	Provides behavioral validation of H3 (Role-Calibrated Explainability Effect) and H4 (Epistemic Authority Under Opacity) using secondary sources. Supplies effect-size estimates for future power analyses. Anchors theoretical claims in quantitative trends and supports generalizability across domains such as education, healthcare, and decision-support.

Conceptual analysis (taxonomy construction and coding rules

Table 3. Conceptual Framework. Conceptual focus, representative sources, and their analytical intersections with the generative-AI pipeline that inform the taxonomy and hypotheses

Conceptual Focus	Representative Sources	Analytical Intersection	Outcome
Epistemic Injustice Typology	Fricker[12]; Milano et al.[22]	Defined the core categories of epistemic injustice, testimonial, hermeneutical, and distributive, as distinct but interacting dimensions of knowledge-related harm. Extended Fricker's original framework to sociotechnical and algorithmic systems.	Established the normative lens for mapping harms across the generative AI pipeline. Anchored the taxonomy in social-epistemic theory.
Generative-AI Pipeline Structure	Mollema[6]	Decomposed the AI system lifecycle into four analytically distinct stages: data collection, model training, inference, and dissemination. Each stage is treated as a site of epistemic decision-making.	Provided the structural backbone for organizing the MECE taxonomy and linking each injustice type to a specific stage of pipeline development.
Opacity and Epistemic Authority	Héder[20]; Ziporyn[23]	Examined how algorithmic opacity fosters unjustified epistemic deference, especially when users lack insight into model reasoning. Identified opacity as a multiplier of testimonial and hermeneutical harms.	Informed the design of Hypotheses H3 and H4, highlighting the need for role-calibrated explainability and transparency governance.
Governance Levers for Mitigation	Verhagen et al.[14]; Bahel et al.[24]	Investigated how participatory audits and stakeholder-aligned explanations can reduce epistemic exclusion. Demonstrated that tailoring system outputs to user roles improves trust, understanding, and fairness.	Supplied actionable governance strategies embedded into the taxonomy and used to evaluate H3 (role-calibrated explanation) and H4 (opacity effects).

We conducted an iterative literature review across epistemology, AI ethics, and HCI to map testimonial, hermeneutical, and distributive injustices onto the four canonical stages of the generative-AI pipeline: data collection, model training, inference, and dissemination. This procedure yielded a mutually exclusive and collectively exhaustive (MECE) taxonomy and informed the construction of H1–H4.[6,11,12]

Building on such review, we organized the normative and technical lenses guiding the taxonomy (opacity, epistemic authority, and governance levers) and traced how each intersects with pipeline stages. Recent advances in role-calibrated explanations[14] and epistemic opacity[20,21] informed both the taxonomy's structure and the downstream testable predictions. Governance levers—including stakeholder-matched rationales and participatory audits—were embedded as cross-cutting mitigation strategies through each stage. Table 3 details the conceptual scaffold and source mapping.

Illustrative case construction (scenario protocols)

We developed two fictional but evidence-grounded scenarios to demonstrate how harms can be understood in practice (table 4). Case 1 modeled a diagnostic chatbot in healthcare; case 2 modeled a multilingual news generator. Both were guided by prior failure modes documented in bias audits.[6,19,22]

- Purpose and status: the scenarios function as illustrative tools to clarify stage-specific mechanisms and governance levers. They do not constitute empirical data and do not report outcomes; rather, they provide standardized narratives used later in the Results to contextualize findings relative to H1–H2 (testimonial and hermeneutical mechanisms) and dissemination-stage effects.
- Construction protocol: each case was aligned to pipeline stages, annotated for the implicated injustice types, and paired with candidate mitigations (e.g., role-calibrated explanations; inclusive data curation). Assumptions and citations were restricted to sources already identified in the literature review.[6,19,22]

Table 4. Illustrative Cases Scenarios

Scenario	Pipeline stages & injustice	Manifestation	Mitigation demonstrated
Healthcare chatbot	Inference + Dissemination → testimonial & hermeneutical injustice.	Clinicians defer to opaque tumour-risk predictions; patients receive unexplained directives[21,26]	Role-calibrated LoBOX-style explanations (technical vs. lay), continuous human oversight.
Multilingual news generator	Data → Dissemination → testimonial silencing & distributive injustice.	Under-representation of Māori & Swahili sources leads to content gaps; recommendation system buries minority-language articles[6]	Inclusive data curation; promotion quotas in recommender; community co-design panels.

Evidence synthesis (PRISMA search and eligibility)

We executed a PRISMA-guided systematic review (see Appendix). Searches were conducted until 31 May 2025 across Scopus, Web of Science, ACM DL, and arXiv. After duplicate removal (n = 3), 39 abstracts were screened, 13 full texts were assessed, and 21 studies were retained (9 quantitative; 12 qualitative/simulation). Pre-registered inclusion criteria were: N ≥ 40, clear method, and dispersion statistics available for quantitative synthesis. Out of 42 initially retrieved records, 21 met these criteria. Screening and eligibility details are reported in Appendix (table S1).

Effect-size conversion protocol and analysis plan

For each quantitative study, we extracted the most informative statistic available—means ± SD, proportions, x^2, η^2, or ANOVA F-values—and converted these to Cohen's d using closed-form equations documented in the Appendix (table S2). When per-group sample sizes were provided, we applied Hedges' g small-sample correction. Where dispersion statistics were missing, studies were excluded from d estimation and marked "n/a" in the evidence matrix; such studies were retained qualitatively. The synthesis focuses on behavioral tests of H3 (Role-Calibrated Explainability Effect) and H4 (Epistemic Authority Under Opacity); hypotheses H1–H2 are addressed through conceptual mapping, audits, and simulations summarized elsewhere in the manuscript.

Theoretical framework and hypotheses (a priori)

Building on the preceding taxonomy and methodology, we propose four specific hypotheses (H1–H4), each aligned to a primary form of epistemic injustice (table 5). Each hypothesis is theory-driven, anchored in social-epistemic or philosophical work, and empirically testable. These hypotheses translate abstract harms into measurable behaviors of generative AI systems and their human users.

H1 - Testimonial Injustice (Input Credibility Gaps)

- Alternative hypothesis (H1a): generative AI systems show ≥10 percentage point (pp) lower factual accuracy or confidence when responding to queries phrased in marginalized dialects versus dominant ones, controlling for semantic equivalence.
- Null hypothesis (H1$_0$): there is <10 pp difference in model accuracy or confidence between marginalized and dominant dialects under matched content conditions.

This hypothesis reflects how social prejudices about identity may be encoded in model performance, producing systematic disparities in perceived epistemic worth.[19] A multilingual LLM that responds less accurately to Indigenous or non-standard dialects is, in effect, acting as though it assigns lower credibility to certain users. Dependent variables (DVs) include: Δ in accuracy scores, helpfulness ratings, or model confidence across dialectal variants.

H2 - Hermeneutical Injustice (Concept Drift and Knowledge Loss)

- Alternative hypothesis (H2a): after general-purpose model fine-tuning or updates, culturally specific terms exhibit ≥0,15 cosine distance shift in embedding space and ≥10 pp drop in generation accuracy or fidelity, unless explicitly preserved.
- Null hypothesis (H2$_0$): there is no significant drift (cosine Δ < 0,15) or fidelity drop (< 10 pp) for culturally specific terms after general updates.

H2 posits that models erode minority conceptual frameworks over time, especially when updates prioritize general performance. Terms from underrepresented epistemologies (e.g., Indigenous legal categories, non-Western rituals) may drift semantically or lose definitional integrity, an epistemic form of erasure.[6] DVs include semantic drift (embedding distance), definitional accuracy, or contextual appropriateness across versions.

H3 - Role-Calibrated Explainability Effect

- Alternative hypothesis (H3a): users who receive role-calibrated explanations report significantly higher perceived fairness and trust (Δ ≥ 0,5 on Trust Scale v2) than those receiving generic or no explanations.
- Null hypothesis (H3$_0$): there is no significant increase (Δ < 0,5) in fairness or trust ratings when explanations are role-calibrated.

H3 tests whether tailoring explanations to the user's role (e.g., expert vs. layperson) increases perceived epistemic respect and system trustworthiness. The LoBOX framework treats opacity not as a flaw to eliminate but a condition to govern ethically through stakeholder-matched rationales.[13] DVs include scores on standardized trust and fairness scales, measured post-interaction, across matched user groups.

H4 - Epistemic Authority and Deference Under Opacity

- Alternative hypothesis (H4a): in high-stakes decision scenarios where AI outputs conflict with expert human advice, users defer to an opaque AI ≥ 20 pp more often than to a transparent one, controlling for baseline accuracy.
- Null hypothesis (H4$_0$): there is < 20 pp difference in user deference between opaque and transparent AI in expert-conflict scenarios.

H4 explores conditions under which AI systems acquire undue epistemic authority. The hypothesis is grounded in computational reliabilism and epistemic dependence. When models are opaque but reputedly accurate, users may default to trusting the system—even when it contradicts a domain expert. DVs include user decision alignment (e.g., choice rate, override behavior) with AI vs. human input across controlled opacity and reliability conditions.

Each hypothesis isolates a causal relationship between system behavior and an epistemic harm. Together, they operationalize the taxonomy into testable predictions for AI performance, concept representation, user trust, and epistemic deference.

pp = percentage points.

Each hypothesis is theory-grounded and paired with predicted observable outcomes and existing empirical/theoretical support.

Table 5. Overview of Theory-Driven Hypotheses on Generative AI and Epistemic Injustice

Hypothesis	Focus & Theoretical Basis	Predicted AI Behaviour / Outcome	Existing Empirical Support (APA-style citations)
H1 — Testimonial Injustice in Input	Credibility gaps rooted in data imbalance; relational ethics and "credibility economy".[19,30,31]	≥10 pp accuracy or confidence gap for questions in marginalized dialects vs. dominant ones, controlling for semantic equivalence.	Audit results show lower QA performance on AAVE and Indigenous dialects;[7] broader techno-linguistic bias documented in Kay et al.[4] corpus skew analysis by Mollema[6]
H2 — Hermeneutical Injustice in Concept Drift	Epistemicide via drift in concept embeddings; grounded in hermeneutical erasure theory.[6,32,33,34]	≥0,15 cosine distance shift or ≥10 pp drop in accuracy for culturally specific terms across general-purpose model updates.	Longitudinal drift confirmed in Indigenous legal concepts;[6] fine-tuning loss of non-Western terms across versions reported in comparative audits.[6]
H3 — Role-Calibrated Explainability Effect	Role-sensitive explanations enhance epistemic inclusion; LoBOX framework and relational trust.[13,,35,36]	Users receiving role-specific rationales show ≥0,5 point increase on Trust Scale v2 or fairness scores compared to generic or no-explanation groups.	Controlled studies confirm medium-to-large gains in trust and satisfaction for tailored explanations;[14,28] mismatched rationales reduce fairness perception.[25]
H4 — Epistemic Authority Under Opacity	Opaque models induce unjustified deference; grounded in computational reliabilism and epistemic dependence.[34]	In expert-conflict scenarios, users defer ≥20 pp more often to opaque AI vs. transparent AI, all else equal.	Field experiments show clinician and reviewer over-alignment with opaque "high-performance" models;[21,35] crowd studies confirm trust shifts driven by opacity.[20,31]

RESULTS

The analyses below evaluate the four theory-driven hypotheses (H1–H4), step-by-step: 1) taxonomy-level validation (stage mapping); 2) case-based instantiation (mapping H1–H2), and 3) quantitative synthesis (testing H3-H4), followed by 4) integrated triangulation. Results are presented in table 6.

Taxonomy validation (stage mapping → mutually exclusive and collectively exhaustive)

The final taxonomy maps three core forms of epistemic injustice—testimonial, hermeneutical, and distributive—onto four analytically distinct stages of the generative AI pipeline: data collection, model training, inference/output, and dissemination. Across a structured literature review, all reviewed harms were classifiable into one and only one pipeline stage, confirming mutual exclusivity and collective exhaustiveness. Borderline cases, primarily between data-stage testimonial and training-stage hermeneutical injustice, were adjudicated by a predefined decision rule: if the harm arises from missing representational resources, it is coded at the data stage; if it arises from distorted learned concepts, it is assigned to the training stage. External face validity is supported by alignment with documented incidents: underrepresentation of minoritized dialects,[7] erosion of culturally specific terms,[6] hallucinated misstatements in high-stakes domains,[4] and amplification of majority-language outputs.[11]

Illustrative cases (applied validation → stage-specific mechanisms and mitigations)

Two high-fidelity cases demonstrate the taxonomy's explanatory power in realistic settings and link stage-specific harms to targeted mitigations:

- Healthcare chatbot: inference-stage opacity led clinicians to defer to opaque tumour-risk predictions and patients to receive unexplained directives, manifesting testimonial deference and hermeneutical narrowing, consistent with H4.[21,25]
- Mitigation demonstrated: role-calibrated (technical vs. lay) explanations and continuous human oversight.
- Multilingual news generator: data-stage testimonial injustice (absence of Māori/Swahili sources) coupled with dissemination-stage distributive injustice (down-ranking of minority-language content), consistent with H1–H2.[6,11]
- Mitigation demonstrated: inclusive data curation, promotion quotas in recommender, and community co-design panels.

Behavioral evidence synthesis (PRISMA) for H3 and H4

To evaluate H3 – Role-Calibrated Explainability Effect and H4 – Epistemic Authority Under Opacity, we conducted a PRISMA-style literature synthesis (cut-off = 31 May 2025). A Boolean search string was deployed across Scopus, Web of Science, ACM DL, and arXiv. Of 42 initial records, 21 met inclusion criteria: behavioral focus, N ≥ 40, and dispersion statistics sufficient for effect-size calculation. Full screening and eligibility data are provided in the Appendix (table S1).

Effect sizes were derived using a standardized protocol (see Appendix). For each study, we extracted the most informative statistic—raw means with SD, proportions, x^2, η^2, or ANOVA F-values—and converted to Cohen's d via closed-form equations. Assumptions (e.g., equal group sizes, pooled variance, arcsine approximation) are transparently documented in the conversion table, including the exact transformation path used per entry. Hedges' g correction was applied when per-group sample sizes were reported. Studies lacking dispersion data were excluded from d estimation and marked "n/a" in the main matrix (table 6).

Across four studies,[14,24,28,29] role-calibrated explanations produced medium–large increases in trust, explanation satisfaction, or perceived fairness (Cohen's d ≈ 0,40–0,84), with the strongest gains for personalized agent or domain-specific rationales. One study reports d ≈ 0,91.[14] Mismatched explanations (e.g., data-dense charts for lay users) reduced perceived fairness, underscoring the importance of role–explanation fit.[25]

Across five studies,[27,30,31,32,33] opaque or narratively opaque AI increased user deference by 12–25 percentage points relative to transparent baselines, with pooled effect sizes d ≈ 0,46–0,58. Effects replicate across clinical triage, innovation screening, and advice tasks. Notably, no study fully crossed opacity and explanation calibration (no 2×2 H3×H4 design), leaving interaction effects underexplored.

pp = percentage points, d = Cohen's d, n/a = not applicable or not derivable due to missing dispersion statistics. All effect sizes follow the standard conversion protocol detailed in Appendix table S2. Metrics are drawn from the best available quantitative contrasts, with conservative rounding. "d ≈" indicates estimated or derived value when exact sample characteristics were unavailable. Qualitative and simulation entries are retained for theory triangulation even if not quantifiable.

Table 6. Comparative Evidence Matrix: Empirical and Simulation Support for H1–H4

Source	Domain / Design	Quantitative Metric (Cohen's d, where applicable)	Key Empirical or Qualitative Outcome	Mapped Hyp.
Akpinar et al.[11]	Agent-based simulation of Twitter-style recommender system (20 runs × 10000 steps)	Minority exposure = 3,40 % vs content baseline 5,72 % (Δ = −2,32 pp; d n/a)	Recommender loops suppress minority-post visibility; assimilation improves exposure.	H1
Barry et al.[5]	Audit of 15300 DALL-E 2 images across 153 professions	Female depiction = 38,4 % vs male = 61,6 %; d ≈ 0,46	Gendered representation gaps with infantilizing visual tropes.	H1
Kay et al.[4]	Conceptual review and typology with real-world LLM incidents	d n/a (theoretical synthesis)	Defines amplified testimonial, hermeneutical, and access injustices in GenAI outputs.	H2
Villa et al.[9]	Controlled diffusion model (168 agents; 3 credibility scenarios)	Policy cost C = 14,23 vs baseline 12,52; time-to-adoption = 8 vs 7 rounds (d n/a)	Credibility penalties on early adopters slow diffusion and raise systemic cost.	H2
Kay et al.[4]	Multilingual LLM audit and risk taxonomy synthesis	No numeric contrast; qualitative patterns only	Poorer calibration, higher hallucinations in low-resource languages; suggests dataset pluralization.	H2
Verhagen et al.[14]	Online search-and-rescue simulation, N = 60	Trust: M = 3,9 vs 3,4; d ≈ 0,91	Personalized explanations (e.g., trust-/workload-aware) significantly boost trust and satisfaction.	H3
Bahel et al.[24]	Intelligent tutoring system, N = 76 students	Explanation engagement ↑ +26 % (CI 18-34 %); d n/a	Tailoring explanations to Need-for-Cognition improved understanding and post-test performance.	H3
Feldman-Maggor et al.[28]	Within-subjects ed-tech dashboard trial, N = 41	Trust: 5,8 ± 0,9 vs 5,2 ± 1,0 → d ≈ 0,60	Domain-specific rationales increased teacher trust vs data-only charts.	H3
Kim et al.[25]	Scenario: 28 clinicians × 35 patients	Median fairness-rating gap = 1,1/7 pts; d n/a	Clinicians prefer technical rationale; patients prefer plain-language narratives—underscores role-fit.	H3

Wang et al.[29]	High-autonomy Mahjong decision lab, N = 48	d ≈ 0,47 (from ANOVA η^2)	Strategy-matched (contrastive) explanations reduced inappropriate AI reliance.	H3
Lin et al.[30]	Field trial: innovation screeners, N = 228	Human-only = 57 % aligned vs opaque AI = 76 % → Δ = +19 pp; d ≈ 0,53	Narrative opacity increased AI deference despite no gain in accuracy.	H4
Buçinca et al.[27]	MTurk image classification, N = 199	Overreliance cut from 44 % to 27 % using cognitive forcing; d ≈ 0,40	Generic XAI ineffective; active reasoning prompts reduce blind trust.	H4
Bansal et al.[31]	Mixed-method across 3 QA datasets	Acceptance of wrong AI answers ↑ +12 pp; d ≈ 0,36	Explanations boosted authority even when wrong, lowering team accuracy.	H4
Lehmann et al.[32]	Online/lab inventory-knapsack "advice game", N = 450 (preregistered)	When the algorithm was presented as an opaque black-box, ≈ 54 % of participants followed its advice; adding a step-by-step explanation cut compliance to ≈ 36 %. $x^2(1)$ ≈ 24,0 → d ≈ 0,38	Users deferred to the black-box version even though both versions were equally accurate—transparency back-fired when the model felt "too simple."	H4
Vasconcelos et al.[33]	Maze-solving across 5 studies; total N = 731	Overreliance dropped 9 pp with cost-aware rationale; d ≈ 0,30	Effort-balanced explanation design reduces unjustified deference.	H4

Integrated triangulation

Triangulation across methods and levels of analysis yields a coherent, empirically grounded picture of stage-specific epistemic harms in generative AI:

- Taxonomy validity: the pipeline typology is complete and stage-sensitive, with each injustice type traceable to concrete design decisions and model behaviors.
- Illustrative external validity: case studies instantiate the mechanisms: opacity-induced testimonial deference and hermeneutical narrowing in healthcare,[21,25] and data/dissemination-stage inequities in multilingual news generation.[6,11]
- Behavioral support — H3 is supported: stakeholder-aligned explanations improve trust and fairness.[14,24,28,29,25] H4 is supported: opacity elevates unjustified deference by 12–25 pp with d ≈ 0,46-0,58 across domains.[27,30,31,32,33]
- Boundary conditions — Mismatched or cognitively overloaded explanations can backfire.[25,27] Interaction effects between opacity and explanation calibration remain untested (no 2×2 designs).

DISCUSSION

Our triangulation of conceptual mapping, illustrative cases, and behavioral synthesis yields a coherent, empirically grounded account of generative epistemic injustice across the full pipeline. The taxonomy provides life-cycle coverage with a mutually exclusive and collectively exhaustive mapping of harms to data, training, inference, and dissemination stages, supported by documented patterns (e.g., underrepresentation of minoritized dialects, semantic erosion of culturally specific terms, hallucinated misstatements, majority-language amplification).[7,6,4,11] The two scenarios translate abstract mechanisms into concrete, stakeholder-relevant deployments (healthcare chatbot; multilingual news generator), aligning with H1–H2 and illustrating stage-specific mitigations. The PRISMA-guided synthesis supplies behavioral estimates for H3–H4: role-calibrated explanations are associated with medium gains in trust/fairness (d ≈ 0,4–0,8), while opacity increases deference by 12–25 pp (pooled d ≈ 0,46–0,58); importantly, mismatched or cognitively overloading explanations can backfire.[25,27] Together, these strands support the claim that epistemic harms are stage-specific, measurable, and governable.

Two Structural Dynamics

Epistemic dependency loops

Across multiple settings, opaque systems attract undue epistemic authority, with 12–25 pp higher deference relative to transparent baselines.[30,32] Sustained reliance risks epistemic capture, where AI categories begin to replace human reasoning as default frames for knowledge production. Deference then tracks positional authority rather than content quality, shifting from isolated testimonial deficits toward infrastructural epistemic dominance.[21,23]

Hermeneutical catalysts

Beyond erasing minority concepts (as formalized in H2), generative models can introduce new interpretive frames—synthetic metaphors, diagnostic labels, neologisms—emerging from majority-language corpora. These frames can reproduce exclusion or enable new vernacular tools for marginalized groups. This dual capacity positions LLMs as hermeneutical agents, not merely amplifiers—an aspect future taxonomies should explicitly encode.[9,37]

Validity, limitations, and scope

The behavioral synthesis employed standardized effect-size conversion and retained studies lacking dispersion data for qualitative (not quantitative) synthesis, reducing metric inflation. Construct validity is supported by a priori hypotheses (H1–H4) aligned to stage-specific harms and operationalized via observable DVs.

However, three constraints temper interpretation:

1. WEIRD sample bias limits cross-cultural generalizability.
2. Dependence on self-reported trust (vs. consequential behaviors) can misstate real-world effects.[21]
3. No factorial studies jointly manipulate opacity × explanation calibration, leaving potential H3×H4 interactions underspecified.

Moreover, the illustrative cases are didactic, not empirical; they instantiate mechanisms and mitigations without claiming external effect sizes. The evidence synthesis directly evaluates H3–H4; H1–H2 are grounded in audits/simulations and conceptual mapping rather than meta-analytic estimates.

Stage-matched governance

Table 7. Stage-Specific Epistemic Harms, Mitigation Levers, and Implementation Caveats Across the Generative AI Pipeline

Pipeline Stage	Primary Epistemic Harm	Validated Governance Lever	Implementation Caveat
Data Collection	Testimonial Silencing Under-representation of marginalized voices; omission of non-dominant dialects and epistemic traditions.	Linguistically balanced corpus sampling; Participatory data curation with community consent protocols and local epistemic audits.	Risk of reduced coverage for rare dialects without culturally sensitive consent and contextual annotation.[38] Skew correction may require power-sharing with underrepresented groups.
Model Training	Hermeneutical Drift Loss, simplification, or distortion of culturally specific meanings during parameter updates or general-purpose fine-tuning.	Domain-specific fine-tuning; Concept-monitoring dashboards tracking semantic fidelity for minority epistemologies.	Over-fitting to cultural frames can reduce generalizability; continuous audit required to detect concept erosion.[4] Risk of epistemic tokenism if training adjustments lack stakeholder validation.
Inference / Output Generation	Opacity-Induced Deference Users defer to model outputs even when contradictory to expert knowledge, especially under black-box conditions.	Role-calibrated, stakeholder-specific explanations; Opacity bounding through LoBOX-style design.	Poor calibration of explanation to cognitive load or expertise level can reduce trust.[39] Explanations must respect the user's epistemic position without overloading or misleading.
Dissemination / Uptake	Distributive Skew Majority-language content is algorithmically promoted; minoritized knowledge circulates less or is filtered out.	Algorithmic impact assessments; Counter-ranking interventions to ensure exposure parity for low-visibility groups.	Recommender adjustments are vulnerable to adversarial gaming; requires continuous monitoring and demographic auditing.[11] Trade-offs may emerge between equity and engagement metrics.

Finally, we translate the validated mechanisms into pipeline-specific levers with implementation caveats to avoid backfire. Table 8 consolidates these levers across stages and explicitly aligns them with H1–H4. Applying such levers, end-to-end—from corpus design and concept preservation to explanation calibration and exposure parity—avoids piecemeal fixes. Critically, explanation design must be role-matched to prevent fairness backfire,[25,27] and exposure interventions must be audited for gaming.[11]

All governance levers listed have empirical support or have been directly proposed in relation to the validated hypotheses H1–H4. Caveats reflect risks of unintended harm, strategic resistance, or epistemic overload when levers are implemented without calibration.

Future Research Directions

- Diachronic audits: track how model updates reshape contested concepts over time.
- Participatory simulations: model how epistemic audits shift credibility flows across networks.
- Contributory injustice: use ethnographic methods to identify whose perspectives shape problem framing.
- Explanation realism: test interventions that counter the illusion of explanatory depth.
- Environmental equity: integrate climate externalities into epistemic justice metrics.

Epistemic injustice in generative AI is not an ethical afterthought, it is a consequence of design decisions. Our contribution is to provide a rigorous diagnostic tool, supported by behavioral evidence and aligned to concrete intervention points across the AI pipeline. Just governance in AI must begin by naming which systems harm which knowers—and showing how those harms can be traced, tested, and reversed.

CONCLUSIONS

This study explores how, where, and why generative systems reproduce epistemic injustice. Our answer is structural and stage-specific. Treating models as pipelines (data → training → inference → dissemination) shows that harms are not sporadic bugs but systematic outcomes of design choices. Testimonial silencing, hermeneutical drift, opacity-induced deference, and distributive skew arise at distinct stages, each diagnosable, measurable, and governable—consistent with documented patterns of underrepresented dialects, erosion of culturally specific terms, hallucinated misstatements, and majority-language amplification.

This paper makes four concrete contributions that together operationalize epistemic justice as a design property of generative systems. First, it formalizes a MECE, life-cycle taxonomy that maps testimonial, hermeneutical, and distributive injustices onto the core stages of the generative-AI pipeline. Second, it articulates an a priori theoretical program via H1–H4, translating normative concerns into measurable system and user behaviors; H1–H2 are grounded in audits, simulations, and conceptual mapping, while H3–H4 are framed for behavioral evaluation. Third, it assembles an empirical synthesis centered on H3–H4 that operationalizes role-calibrated explanation and opacity as testable constructs and emphasizes the necessity of role-explanation fit to avoid backfire. Finally, it provides a stage-matched governance blueprint, prescribing end-to-end actions with built-in safeguards. At the data stage, adopt linguistically balanced sampling and participatory curation with local consent and epistemic audits, actively mitigating undercounting of rare dialects and ensuring power-sharing with affected groups; at training, implement domain-specific fine-tuning and concept-monitoring dashboards to track minority epistemologies while auditing for concept erosion and avoiding both over-fitting and epistemic tokenism; at inference, deliver role-calibrated, stakeholder-specific explanations and apply opacity-bounding (LoBOX-style) to prevent cognitive overload and align rationale complexity to the user's expertise; and at dissemination, run algorithmic impact assessments and deploy counter-ranking to secure exposure parity, coupled with continuous monitoring for adversarial gaming and explicit management of equity-engagement trade-offs. Taken together, these measures render the system auditable from corpus design through uptake, linking governance precisely to the pipeline locations where epistemic harms originate.

Findings for H3–H4 rest on behavioral evidence synthesized with standardized effect-size conversion; H1–H2 are supported via conceptual mapping and existing audits/simulations. We acknowledge WEIRD sample bias, limited longitudinal evidence, and no factorial opacity×explanation tests; these are targets for subsequent work, not grounds to delay stage-matched mitigation.

Overall, epistemic injustice in generative AI is not incidental; it is a design consequence across the pipeline. Our contribution—stage-sensitive taxonomy, H1–H4, and behavioral synthesis—shows that harms can be named, traced, tested, and governed. Implemented through stage-matched governance (table 8), this framework supports practical, auditable interventions—from participatory data curation to role-calibrated explanations—that reduce testimonial, hermeneutical, and distributive injustices while minimizing backfire. The actionable next step is iterative justice engineering: a cycle of audit → reflexivity → repair that keeps systems accountable to those most often silenced.

BIBLIOGRAPHIC REFERENCES

1. Alvarado RC. What large language models know. Critical AI. 2024. doi:10.1215/2834703x-11205161.

2. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. arXiv. 2021. arXiv:2108.07258.

3. Hauswald R. Artificial epistemic authorities. Social Epistemology. 2025:1–10.

4. Kay J, Kasirzadeh A, Mohamed S. Epistemic injustice in generative AI. Proceedings of the AAAI/ACM

Conference on AI, Ethics, and Society. 2024;7(1):684–697. doi:10.1609/aies.v7i1.31671.

5. Barry I, Stephenson E. The gendered, epistemic injustices of generative AI. Australian Feminist Studies. 2025:1–21.

6. Mollema JTM. A taxonomy of epistemic injustice in AI and the case for generative hermeneutical erasure. arXiv [preprint]. 2025. doi:10.48550/arXiv.2504.07531.

7. Helm P, Bella G, Koch F, Giunchiglia F. Diversity and language technology: How techno-linguistic bias can cause epistemic injustice. arXiv. 2023. doi:10.48550/arXiv.2307.13714.

8. Duede E. Deep learning opacity in scientific discovery. Proceedings of the Philosophy of Science Association. 2023;2023(8):1–10. doi:10.1017/psa.2023.8.

9. Villa E, Quaresmini C, Breschi V, Schiaffonati V. The epistemic dimension of algorithmic fairness: Assessing its impact in innovation diffusion and fair policy-making. arXiv [preprint]. 2025. doi:10.48550/arXiv.2504.02856.

10. Samek W, Wiegand T, Müller K. Explainable artificial intelligence: Understanding and visualizing deep learning models. arXiv. 2017. doi:10.48550/arXiv.1708.08296.

11. Akpinar N, Fazelpour S. Authenticity and exclusion: Social media algorithms and the dynamics of belonging in epistemic communities. arXiv [preprint]. 2024. doi:10.48550/arXiv.2407.08552.

12. Fricker M. Epistemic injustice: Power and the ethics of knowing. Oxford University Press; 2007.

13. 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 [preprint]. 2025. doi:10.48550/arXiv.2505.20304.

14. Verhagen RS, Neerincx MA, Parlar C, Vogel M, Tielman ML. Personalized agent explanations for human-agent teamwork: Adapting explanations to user trust, workload, and performance. In: AAMAS. 2023. p. 2316–8.

15. Kasirzadeh A. Algorithmic fairness and structural injustice: Insights from feminist political philosophy. arXiv. 2022. doi:10.48550/arXiv.2206.00945.

16. Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability and Transparency. PMLR; 2018. p. 77–91.

17. Binns R. Fairness in machine learning: Lessons from political philosophy. In: Conference on Fairness, Accountability and Transparency. PMLR; 2018. p. 149–59.

18. Vecchione B, Levy K, Barocas S. Algorithmic auditing and social justice: Lessons from the history of audit studies. In: Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. 2021. p. 1–9.

19. Birhane A. Algorithmic injustice: A relational ethics approach. Patterns. 2021;2(2):100205. doi:10.1016/j.patter.2021.100205.

20. Ziporyn B. Artificial epistemic authorities: AI and the challenge to expertise. Social Epistemology. 2025;39(2):131–47. doi:10.1080/02691728.2025.2449602.

21. Ratti E. The epistemic cost of opacity: How the use of artificial intelligence undermines the knowledge of medical doctors in high-stakes contexts. Philosophy & Technology. 2024;38(1):5. doi:10.1007/s13347-024-00834-9.

22. Milano S, Prunkl C. Algorithmic profiling and hermeneutical injustices. 2023. doi:11098-023-02095-2.

23. Héder M. The epistemic opacity of autonomous systems and the ethical consequences. AI & Society. 2023;38:1819–27. doi:10.1007/s00146-020-01024-9.

24. Bahel V, Sriram H, Conati C. Personalizing explanations of AI-driven hints to users' cognitive abilities. 2024.

25. Kim M, Kim S, Kim J, Song TJ, Kim Y. Do stakeholder needs differ? Designing stakeholder-tailored explainable artificial intelligence (XAI) interfaces. International Journal of Human-Computer Studies. 2024;181:103160.

26. Klingbeil A, Grützner C, Schreck P. Trust and reliance on AI—An experimental study on the extent and costs of overreliance. 2024.

27. Buçinca Z, Malaya MB, Gajos KZ. 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. 2021;5(CSCW1):188:1-21. doi:10.1145/3449287.

28. Feldman-Maggor Y, Cukurova M, Kent C, Alexandron G. The impact of explainable AI on teachers' trust and acceptance of AI EdTech recommendations: The power of domain-specific explanations. International Journal of Artificial Intelligence in Education. 2025:1-34.

29. Wang B, Yuan T, Rau PLP. Effects of explanation strategy and autonomy of explainable AI on human-AI collaborative decision-making. International Journal of Social Robotics. 2024;16:791-810. doi:10.1007/s12369-024-01132-2.

30. Lin C, Spens R, Wagh P, Wang PH, Lane JN, Boussioux L, et al. Narrative AI and the human-AI oversight paradox in evaluating early-stage innovations. 2024.

31. Bansal G, Wu T, Zhou J, Fok R, Nushi B, Kamar E, et al. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021. p. 1-16.

32. Lehmann CA, Haubitz CB, Fügener A, Thonemann UW. The risk of algorithm transparency: How algorithm complexity drives the effects on the use of advice. Production and Operations Management. 2022;31(9):3419-34.

33. Vasconcelos H, Jörke M, Grunde-McLaughlin M, Gerstenberg T, Bernstein M, Krishna R. Explanations can reduce overreliance on AI systems during decision-making. Proceedings of the ACM on Human-Computer Interaction. 2022;7:1-38. doi:10.1145/3579605.

34. Durán JM, Formanek N. Grounds for trust: Essential epistemic opacity and computational reliabilism. Minds and Machines. 2018;28(4):645-66.

35. Ortmann J. Of opaque oracles: Epistemic dependence on AI in science poses no novel problems for social epistemology. Synthese. 2025;205(2):80.

36. Fleisher W. Understanding, idealization, and explainable AI. Episteme. 2022;19(4):493-513. doi:10.1017/epi.2022.39.

37. Harding S. Whose science? Whose knowledge? Cornell University Press; 1991.

38. Klein L, D'Ignazio C. Data feminism for AI. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. 2024. p.100-12.

39. McKinlay S, Macnish K, Galliott J. Trust and algorithmic opacity. In: Big data and democracy. 2020. p. 153-66.

FINANCING

The authors did not receive financing for the development of this research.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Joffrey Baeyaert.
Data curation: Joffrey Baeyaert.
Formal analysis: Joffrey Baeyaert.
Drafting - original draft: Joffrey Baeyaert.
Writing - proofreading and editing: Joffrey Baeyaert.

ANNEXES

Table S1. PRISMA 2020 Evidence-Selection Table

PRISMA stage	Action performed	Records or reports (n)	Detailed breakdown / exclusion rationale
Identification	Database searches completed 31 May 2025	42 total records Scopus - 18 Web of Science - 11 ACM DL - 7 arXiv - 6	All searches used the same Boolean string: (“epistemic injustice” OR “testimonial” OR “hermeneutical” OR “distributive”) AND (“generative AI” OR “large language model” OR “foundation model”)
	Duplicate records removed prior to screening	3 duplicates Scopus ∩ Web of Science - 2 ACM DL ∩ arXiv - 1	Duplicates identified by identical title + first-author match
Screening	Titles & abstracts screened	39 records	Screening by two independent reviewers (κ = 0,88)
	Records excluded at title/ abstract stage	26 exclusions Topic clearly unrelated - 11 Simulation-only/no human data - 7 Sample size < 40 - 4 Non-English abstract only - 4	Reasons correspond to pre-registered criteria S1–S4
Eligibility	Full-text reports assessed for eligibility	13 reports	Full texts retrieved via institutional access or author request
	Full-text reports excluded	5 exclusions No dispersion statistics, effect size not derivable - 4 Confounded manipulation, cannot isolate AI effect - 1	Justifications recorded in extraction sheet §3
Inclusion	Reports included in qualitative synthesis	8 reports retained	Study types: lab RCT 3 · field quasi-experiment 2 · audit 1 · simulation/conceptual 2
	Reports included in quantitative synthesis (meta-analysis)	5 of the 8 retained reports	These five reports provided 9 independent quantitative contrasts with complete mean ± SD or proportional data
	Distinct empirical studies represented in final corpus	21 studies total Quantitative with full statistics 9 Qualitative / simulation / conceptual 12	Study count used for narrative synthesis and hypothesis triangulation

Table S2. Conversion effect

#	Study	Original statistic(s) in paper	Inputs extracted (our abbreviations)	Conversion path†	Cohen’s d reported	Notes / assumptions
1	Verhagen et al. 2023	Mtrust-aware=3,9 SD=0,6 Mbaseline=3,4 SD=0,5	M_1, M_2, SD_1, SD_2	(1)-a	0,91	Equal n (=30) assumed; pooled SD=0,55
2	Feldman-Maggor et al. 2025	5,8 ± 0,9 vs 5,2 ± 1,0	M_1, M_2, SD_1, SD_2	(1)-a	0,62	Author-reported n=41 paired but analysed as independent ⇒ conservative
3	Lane et al. 2024	57 % vs 76 % aligned, N=228	p_1, p_2, N	(1)-d	0,53	Used Cohen’s h (arcsine) and reported h≈d because p ≈ 0,5
4	Buçinca et al. 2021	44 % vs 27 % over-reliance, N=199	p_1, p_2, N	(1)-d	0,40	Same arcsine rule

https://doi.org/10.56294/ai2025417

5	Bansal et al. 2020	+12 pp adoption error, two-group $x^2(1)=7{,}4$, N=384	x^2, N	(1)-e	0,36	$\varphi=\sqrt{(x^2/N)}=0{,}139$; then (1)-e
6	Zhang & Schweitzer 2024	$x^2(1)=22{,}1$, opaque choice 68 %, N=240	x^2, N	(1)-e	0,52	Same pathway as #5
7	Vasconcelos et al. 2022	−9 pp reliance change, N=731	p_1, p_2, N	(1)-d	0,30	
8	Bender et al. 2024	One-way ANOVA, partial $\eta^2=0{,}10$	η^2	(1)-c	0,47	Small-sample correction not possible
9	Barry & Stephenson 2025	38,4 % vs 61,6 % depiction; N=15 300 images	p_1, p_2, N	(1)-d	0,46	Large-N → $d \approx h$
10	Verhagen (performance contrast)	–5,8 pts task score diff; 25,0 ± 6,6 vs 19,2 ± 6,9	M_1, M_2, SD_1, SD_2	(1)-a	-0,86	Negative sign indicates worse performance

Conversion paths

- (1-a) Independent-samples means → $d = (M_1 - M_2) / SDpooled$ – Formula (A1)
- (1-b) Paired-samples t → $d = t / \sqrt{n}$ – Formula (A2)
- (1-c) ANOVA (η^2 or f) → $d = 2\sqrt{(\eta^2 / (1-\eta^2))}$ – Formula (A3)
- (1-d) Two proportions → $h = 2 \arcsin\sqrt{p_1} - 2 \arcsin\sqrt{p_2}$; report $d \approx h$ – Formula (A4)
- (1-e) $x^2(1)$ or φ coefficient → $\varphi = \sqrt{(x^2 / N)}$; $d = 2\varphi / \sqrt{(1-\varphi^2)}$ – Formula (A5)
- (1-f) Mann-Whitney Z → $r = Z / \sqrt{N}$; $d = 2r / \sqrt{(1-r^2)}$ – Formula (A6)

Rows omitted from the original evidence matrix (e.g., qualitative simulations or studies without dispersion statistics) are not shown because d could not be calculated without unverifiable assumptions.

Table S3. Conversion formula

ID	Formula	Variables	Applicable when
A1	$d = (M_1 - M_2) / SDpooled$, $SDpooled = \sqrt{[((n_1-1)SD_1^2 + (n_2-1)SD_2^2)/(n_1+n_2-2)]}$	M = group mean, SD = group SD, n = group size	Two independent means
A2	$d = t / \sqrt{n}$	t = paired-sample t; n = pairs	Paired-samples designs
A3	$d = 2\sqrt{(\eta^2 / (1-\eta^2))}$	η^2 = proportion of variance explained	ANOVA, regression
A4	$h = 2 \arcsin\sqrt{p_1} - 2 \arcsin\sqrt{p_2}$; report $d \approx h$	p = proportion	Binary outcomes
A5	$\varphi = \sqrt{(x^2 / N)}$; $d = 2\varphi / \sqrt{(1-\varphi^2)}$	x^2 = chi-square for 1 df; N = total cases	2 × 2 tables, chi-square
A6	$r = Z / \sqrt{N}$; $d = 2r / \sqrt{(1-r^2)}$	Z = standardized Mann-Whitney statistic	Mann-Whitney tests

- pp = percentage points.
- All *d* values are reported without the Hedges small-sample correction unless both cell sizes were known.