

Correcting AI-generated false criminal allegations: a human-rights framework for remedial closure

Hanane Merzougui

Southeast University

Xianghe Gong

gongxianghe07@outlook.com

Southeast University

Yaohui Liu

Southeast University

Article

Keywords: artificial intelligence, data protection, effective remedy, human rights, rectification, algorithmic accountability

Posted Date: October 18th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11036486/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Generative artificial intelligence (AI) can fabricate criminal allegations about identifiable people, but the legal difficulty is not confined to correcting the visible output. A false claim may persist in service-side memory, retrieval systems, application programming interface transmissions, risk fields and downstream institutional decisions. Existing scholarship on data rectification, AI supply chains, algorithmic contestability and model correction addresses parts of this problem but does not provide a unified legal test for when distributed remediation is complete. This article develops a propagation-aware standard of remedial closure: a structured test for determining when corrective action has been completed across relevant stages of an AI system. It combines doctrinal analysis of international and regional human-rights law, European Union data-protection law and business-and-human-rights standards with technical literature on memory, retrieval, model editing and unlearning. China's Personal Information Protection Law is used as a limited comparative portability test, and a publicly documented complaint concerning an AI-generated false criminal allegation provides a case-informed scenario. The framework requires five functions: accessible challenge and reasoned determination; cessation at reasonably identifiable nodes; correction of inaccurate records and reassessment of materially affected decisions where applicable law permits or requires; verifiable implementation; and proportionate non-recurrence safeguards. Responsibility remains actor- and operation-specific, and tracing may cease where further inquiry is objectively impossible or disproportionate considering severity and expected remedial value. Rectification is treated as legally complete only when the responsible actor can provide reasonable, demonstrable assurance that the verified false allegation has been corrected or contained across reasonably identifiable, materially relevant processing nodes within its authority, with applicable downstream effects addressed and residual risks documented.

1. Introduction

A fabricated criminal allegation attributed to an identifiable person can impair privacy, dignity, reputation and social participation, especially when it influences consequential decisions. The harm does not end when a chatbot displays the allegation: the same proposition may persist through memory, retrieval, API transmission or downstream institutional use. Such persistence can transform informational harm into institutional harm (Citron and Solove, 2022). Here, 'infrastructural' refers to the embedding of false criminal status in records or workflows.

The problem is illustrated by the complaint filed in March 2025 by the European Center for Digital Rights (Noyb) on behalf of Arve Hjalmar Holmen. The complaint alleged that ChatGPT combined accurate biographical information with an invented claim that Holmen had murdered two of his children, attempted to murder a third and received a twenty-one-year prison sentence (Noyb, 2025). This study does not determine the provider's liability in that dispute. It uses the publicly documented incident to identify a remedial problem: correcting one displayed answer may not eliminate recurrence where inaccurate information persists in retrieval systems or agent memory, because stored errors can influence and propagate into later outputs (Mallen et al., 2023; Xiong et al., 2026).

European data-protection law supplies a clear starting point. Where it is processed as factual information about an identifiable living person, a fabricated assertion that the person has been convicted and imprisoned constitutes inaccurate personal data for the purposes of the GDPR's accuracy and rectification framework. The General Data Protection Regulation (GDPR) requires accuracy, transparent procedures and rectification without undue delay (GDPR, 2016, arts. 5(1)(d), 12, 16). The Court of Justice of the European Union (CJEU) has linked rectification to Article 8(2) of the Charter of Fundamental Rights of the European Union (EU Charter) and accepted that a claimant may be required to provide relevant evidence of inaccuracy (Deldits, 2025, paras. 24, 41), while *Nowak* confirms that accuracy must be assessed in light of the purposes for which the data are processed (Nowak, 2017, para. 53). The unresolved question arises after falsity is established: how far must corrective action extend before the remedy is legally complete?

Technical research identifies factual inaccuracies in language-generation systems and limitations in output-level factuality assessment (Ji et al., 2023; Bender et al., 2021; Lin et al., 2022; Min et al., 2023; Manakul et al., 2023). Existing scholarship examines rectification, AI-generated reputational harm, distributed responsibility, algorithmic contestability and technical correction (Häuselmann and Custers, 2024; Herrerias Castro, 2026; Graßhoff et al., 2026; Kinchin, 2024; Weidinger et al., 2022). However, the literature examined does not establish an integrated legal standard for determining when distributed remediation can be considered complete.

The research question is: once an AI-generated false criminal allegation has moved beyond the original output, what reviewable conditions establish completion of an effective remedy? This article develops a propagation-aware rectification standard incorporating five functions: accessible challenge and reasoned determination; cessation at reasonably identifiable nodes; correction of inaccurate records and reassessment of materially affected decisions where applicable law permits or requires; verifiable implementation; and proportionate safeguards against recurrence. Its contribution is a bounded test of remedial closure linking legal duties to technical propagation, allocating responsibility by processing operation and establishing a principled stopping rule.

2. Methods and analytical design

The analysis is doctrinal and interdisciplinary, with a limited comparative portability test. It prioritises international and regional human-rights law; the EU Charter, GDPR, CJEU and EDPB materials; the AI Act; and the UNGPs. Technical literature was selected for relevance to factuality, memory, retrieval, model editing, unlearning, auditability and agentic workflows. The Holmen complaint is a case-informed starting point, and the six-node chain is an analytical abstraction rather than evidence of proprietary architecture.

China is used as a limited comparative portability test, not as a claim of institutional equivalence. Its PIPL, PRC Civil Code and specialised AI rules provide a materially different architecture in which to test whether the framework's core functions can be expressed through alternative legal mechanisms.

Functional portability despite different supervisory and enforcement structures would reduce dependence on the GDPR's institutional design. Legal and technical developments were checked through 27 August 2026.

Relevant legal and interdisciplinary scholarship was identified through targeted searches of scholarly and official repositories, supplemented by citation chaining, using terms concerning AI hallucination, inaccurate personal data, rectification, model editing, unlearning, contestability, effective remedy and AI supply chains.

Source selection followed a structured purposive protocol. Legal authorities were selected for their relevance to accuracy, rectification, recipient communication, effective remedy, responsibility and automated decisions. Technical studies were selected for their relevance to persistence, correction, provenance, auditability and recurrence across the six-node scenario. Primary legal texts were preferred, and the selected materials were mapped against the six nodes to assess the legal and operational basis of each remedial step. This structured analysis is not a systematic review and does not estimate prevalence, effect size or comparative cost.

The analysis distinguishes express legal duties, interpretive applications and proposed normative safeguards. Express duties are grounded in enacted law or binding case law, including accuracy, access, rectification and recipient communication (GDPR, 2016, arts. 5(1)(d), 15, 16, 19). Interpretive applications specify how such duties may operate in distributed AI systems. Normative safeguards, including the proposed correction-provenance record and elements of recurrence testing, are presented as institutional or evidentiary proposals rather than freestanding legal rights. This distinction avoids converting technical best practice or soft law into binding doctrine without authority.

3. Human-rights foundations and operational law

3.1 From fabricated criminal allegation to infrastructural criminal identity

A fabricated criminal allegation becomes an infrastructural criminal identity when embedded in records, classifications or workflows that influence institutional treatment of the affected person. Automated systems may transform inaccurate information into scores or adverse determinations before the person discovers the record (Crawford and Schultz, 2014; Citron, 2008). The concept describes an operational harm rather than a new legal status. Here, 'infrastructural criminal identity' denotes the harm, 'propagation-aware rectification' the remedial process, and 'remedial closure' its legal endpoint.

Article 8 of the European Convention on Human Rights (ECHR), read with Article 13, provides an important foundation for protection against persistent false criminal-status information (ECHR, 1950, arts. 8, 13). The ECtHR has recognised that retention, disclosure and searchability of personal information may engage private life, particularly where inaccurate or sensitive information produces

continuing consequences (Rotaru v Romania, 2000; Khelili v Switzerland, 2011; M.M. v United Kingdom, 2012). Digital-archive cases further show that remedies must balance reputational protection against freedom of expression and historical accuracy (Hurbain v Belgium, 2023; M.L. and W.W. v Germany, 2018).

Digital-archive cases illuminate proportionality. In *Hurbain v Belgium*, anonymisation of an old article was accepted where its continued unrestricted online availability was apt to create a ‘virtual criminal record’ and cause serious reputational harm (Hurbain v Belgium, 2023, paras. 232–256). *M.L. and W.W. v Germany* upheld, following an Article 8–Article 10 balancing exercise, the continued online availability of accurate reporting of genuine criminal proceedings (M.L. and W.W. v Germany, 2018, paras. 89–90, 97–105, 111–116), and *Biancardi v Italy* treated de-indexing as a narrower alternative to deletion (Biancardi v Italy, 2021). These authorities concern accurate journalism rather than AI-fabricated criminal allegations, but they establish two relevant propositions. First, persistence and searchability can intensify informational harm. Secondly, a remedy must be accurately targeted: it should neutralise the false status without suppressing accurate reporting, criticism or the historical fact that an allegation was generated.

3.2 The content of an effective remedy

The right to an effective remedy links substantive protection to practical correction. The EU Charter protects private life, personal data and judicial redress (EU Charter, 2016, arts. 7, 8, 47). The ICCPR protects privacy and requires effective remedies, while General Comment No 31 emphasises cessation and adequate reparation (ICCPR, 1966, arts. 2(3), 17; Human Rights Committee, 2004, paras. 15–16). Comparable guarantees appear in the American and African human-rights systems (American Convention on Human Rights, 1969; African Charter on Human and Peoples’ Rights, 1981). The UN Basic Principles identify cessation, restitution, compensation, rehabilitation, satisfaction and non-repetition as remedial measures; they are used here by analogy rather than as directly binding rules for private AI disputes (UN General Assembly, 2005, principles 18–23). The Council of Europe AI Convention also addresses complaints, documentation and procedural safeguards, although it had not entered into force by the study’s cut-off date (Council of Europe, 2024, arts. 14–15). From these sources, the study derives five remedial functions (Human Rights Committee, 2004, paras. 15–17; UN General Assembly, 2005, principles 11–24; Council of Europe, 2024).

1. **Access and determination.** Under the proposed framework, the person should be able to identify the disputed proposition, submit evidence reasonably available to them, obtain competent human review and receive a reasoned determination on accuracy.
2. **Cessation.** Under the proposed framework, once falsity is verified, corrective action should prevent the proposition from continuing to operate at reasonably identifiable nodes within the actor’s responsibility.

3. **Restoration.** Under the proposed framework, inaccurate records should be corrected, and materially affected assessments or decisions should be reconsidered where applicable law permits or requires.
4. **Implementation and verification.** Under the proposed framework, corrective steps should be documented in a form capable of supervisory or judicial inspection.
5. **Non-recurrence and reviewable closure.** Proportionate testing should address reappearance through paraphrase, memory, retrieval or system updates, and the remedial process should permit renewed review if material recurrence is later detected.

Explanation alone cannot correct inaccurate records or reverse their consequences. Scholarship treats explanation as facilitating contestation rather than exhausting algorithmic accountability (Edwards and Veale, 2017; Kaminski, 2019). In *Dun & Bradstreet Austria*, the CJEU required an intelligible explanation of the procedure and principles applied, with protected information supplied to the competent supervisory authority or court where necessary for balancing (Dun & Bradstreet Austria, 2025, paras. 58–76). Counterfactual explanations may help claimants identify influential inputs and contest adverse decisions (Wachter et al., 2018), but they do not establish that inaccurate data were corrected or that downstream decisions were reconsidered.

3.3 State duties, business responsibility and EU implementation

Three layers of responsibility should remain distinct. States are bound by treaty duties to respect rights, protect against serious private interference and provide effective remedies (Human Rights Committee, 2004, paras. 3, 8, 15). Businesses have a UNGP responsibility to avoid causing or contributing to adverse human-rights impacts and to provide for or cooperate in remediation where they have caused or contributed to them (OHCHR, 2011, principles 13, 15, 22). Domestic and regional law, by contrast, creates directly enforceable duties for controllers and processors. UNGP Principle 31 requires effective non-judicial grievance mechanisms to be legitimate, accessible, predictable, equitable, transparent, rights-compatible and a source of continuous learning, while operational-level mechanisms should additionally be based on engagement and dialogue (OHCHR, 2011, principle 31). Such procedures can provide a first route to correction, but independent supervisory or judicial review should remain available where applicable law provides it. Human-rights materials therefore identify the remedial functions against which effectiveness is assessed; binding duties imposed directly on private providers must derive from the applicable legally operative regime.

Risk-management standards support implementation without creating directly enforceable individual rights. NIST, UNESCO and OECD frameworks emphasise lifecycle governance, documentation and monitoring (NIST, 2023; NIST, 2024; UNESCO, 2021; OECD, 2024). Scholarship similarly connects lifecycle accountability, traceability and technical verification with effective oversight (McGregor et al., 2019; Council of Europe Committee of Ministers, 2020; Cobbe et al., 2021; Kroll et al., 2017). Regulators may

therefore need powers to preserve logs, obtain data-flow and recipient records, order restriction or rectification, and scrutinise claims of impossibility or disproportionate effort.

Within its scope, the GDPR operationalises the five functions. Article 5 requires accuracy and accountability; Article 12 accessible procedures; Articles 16 and 18 rectification and restriction; Article 19 communication to recipients subject to impossibility or disproportionate effort; Articles 24 and 25 organisational and design measures; and Articles 77–82 supervisory, judicial and compensatory routes (GDPR, 2016). Access supports correction by revealing processed data. CRIF requires a faithful and intelligible reproduction of personal data and, where essential to exercising rights, relevant document or database extracts (CRIF, 2023); Pankki S extends access to recorded consultation dates and purposes (Pankki S, 2023, paras. 62, 69–70, 83); and EDPB guidance supports layered, readable disclosure (EDPB, 2023, paras. 139–147).

Under the proposed framework, evidential allocation should reflect informational asymmetry. In the CJEU's inaccurate-content judgment, a claimant was not required to obtain a prior judicial declaration of falsity but had to provide evidence capable of substantiating manifest inaccuracy (Google (Dereferencing of Allegedly Inaccurate Content), 2022, paras. 68, 72). *Deldits* similarly permits relevant and sufficient evidence that may reasonably be required of the data subject to establish inaccuracy (Deldits, 2025, para. 41). Under the proposed framework, once the claimant identifies the precise disputed proposition and supplies reasonably obtainable evidence supporting *prima facie* falsity, the controller should investigate relevant non-public processing nodes within its responsibility. Where inaccurate data influence a consequential automated decision, *SCHUFA* confirms that Article 22 may apply if the score effectively determines a third party's contractual decision (*SCHUFA Holding*, 2023, paras. 40–50, 73). Guidance on automated decisions requires meaningful rather than nominal human intervention (Article 29 Data Protection Working Party, 2018), while early commentary on the draft legislation identified unresolved allocation-of-responsibility problems and warned that framing transparency as the main response could legitimise problematic practices without resolving deeper substantive concerns (Veale and Zuiderveen Borgesius, 2021).

The AI Act establishes logging and human-oversight requirements for high-risk AI systems, a right to lodge complaints concerning infringements of the Regulation, and a right to explanation for specified individual decisions based on certain Annex III high-risk systems (EU AI Act, 2024, arts. 12, 14, 85–86). Regulation (EU) 2026/1744 deferred the application of Chapter III, Sections 1–3, with the exception of Article 6(5), to 2 December 2027 for systems classified as high-risk under Article 6(2) and Annex III and to 2 August 2028 for systems classified as high-risk under Article 6(1) and Annex I. Accordingly, some high-risk-system mechanisms relied upon here are enacted but remained prospective at the study's cut-off date; the amendment did not displace GDPR accuracy and rectification duties (Digital Omnibus on AI, 2026; GDPR, 2016, arts. 5(1)(d), 16). EDPB guidance on data protection by design supports a system-level, lifecycle approach in which safeguards remain effective throughout processing, data-subject rights are facilitated, accuracy is maintained and accumulated errors are mitigated across the processing chain (EDPB, 2020, paras. 13–38, 65, 77–79). The ChatGPT Taskforce further states that responsibility

for GDPR compliance cannot be shifted to data subjects and that transparency about ChatGPT's probabilistic and potentially inaccurate outputs is not sufficient, by itself, to satisfy the accuracy principle (EDPB, 2024b, paras. 24–31). Opinion 28/2024 complements this approach through case-specific assessment, safeguards, documentation, testing and corrective measures across the development and deployment of AI models (EDPB, 2024a, paras. 34–58, 96–116, 127–132).

A limited comparison with China tests whether the five functions remain intelligible outside the EU institutional structure. The PIPL provides rights and handler duties concerning access, correction or supplementation, specified deletion, automated decision-making, organisational safeguards, impact assessment, audits and supervision (PIPL, 2021, arts. 24, 44–50, 51, 54–58, 60–66). The PRC Civil Code protects human dignity, reputation, privacy and personal information and recognises remedies including cessation of infringement, correction or deletion of inaccurate content or personal information, elimination of adverse effects and rehabilitation of reputation (PRC Civil Code, 2020, arts. 990–991, 995, 1024, 1028–1029, 1032–1038). China's compliance-audit and network-data rules support verification through complaint handling, rights-exercise channels, periodic compliance audits and supervisory inspection (Cyberspace Administration of China, 2025; Regulations on Network Data Security Management, 2024, arts. 20, 23, 27, 50), while rules on algorithmic recommendation and deep synthesis add duties relevant to content governance, complaints and verification; the specialised generative-AI rules additionally address the accuracy and reliability of generated content (Cyberspace Administration of China et al., 2021; Cyberspace Administration of China et al., 2022; Generative AI Interim Measures, 2023). The Supreme People's Court's facial-recognition interpretation further illustrates civil personality-rights remedies (Supreme People's Court, 2021, arts. 2–3, 6, 8–12). This comparison shows direct functional portability for access, correction and cessation, and functional substitutes for verification through audit and compliance mechanisms. It does not establish identical rules on recipient notification, evidential allocation or independent review. China therefore operates as a stress test of the framework's functions, not as institutional equivalence: the function may travel, while the legal route, responsible institution and available remedy remain system-specific.

4. Propagation, responsibility and evidentiary asymmetry

4.1 The six-node chain

Agentic systems, meaning AI systems capable of pursuing tasks through memory, tools and multi-step actions, may retrieve information, invoke tools, exchange data through APIs and influence later workflows (Beduschi, 2026; Goldstein, 2025). For remedial analysis, this article identifies six relevant nodes: generated output; service-side memory; retrieval index; API disclosure; downstream risk field (such as a classification, score or internal assessment derived from the false information); and consequential institutional decision. These nodes are legal-analytical categories rather than claims about any particular technical architecture.

A propagation edge is an evidenced or technically plausible transfer or semantic derivation through which the false proposition moves between processing nodes. Under the proposed framework, propagation-aware rectification is limited to reasonably identifiable nodes within an actor's responsibility where the false proposition is stored, transmitted, transformed or materially relied upon. It does not extend to every conceivable future disclosure.

The visual boundary between originating and downstream actors is heuristic rather than jurisdictional or categorical: a single actor may control operations on both sides of the boundary, while different actors may share or successively exercise legally relevant influence over the same processing pathway.

Figure 1 presents a schematic and non-exhaustive propagation model for remedial analysis. The arrows indicate plausible functional pathways rather than a necessary processing sequence or an empirical reconstruction of any particular provider's architecture. It visualises the six-node propagation chain, the illustrative allocation boundary between originating and downstream actors, and the distinction between output correction, proposition-level rectification and propagation-aware rectification.

The widening remedial scope illustrated in Fig. 1 does not imply unlimited tracing. The inquiry remains bounded by responsibility, reasonable identifiability, materiality and proportionality, and remedial closure may coexist with documented residual risks outside the responsible actor's authority.

4.2 Operation-specific responsibility

Propagation does not create universal chain-wide liability. Under GDPR Article 4(7), controllership turns on who determines the purposes and means of the relevant processing operation. The CJEU applies a functional approach to controllership, examining whether an actor contributes to determining the purposes and means of processing, as *Wirtschaftsakademie* (2018, paras. 27–43) and *Jehovan todistajat* (2018, paras. 65–75) demonstrate. Google Spain confirms that a search-engine operator may be a controller for its own organisation and presentation of personal data even though the underlying information was published by third parties (Google Spain, 2014, paras. 33–41). At the same time, *Fashion ID* limits joint responsibility to operations over which the actor exercises relevant influence (Fashion ID, 2019, paras. 68–85), and IAB Europe confirms that joint controllership does not automatically extend to subsequent processing for which the actor does not determine the purposes or means (IAB Europe, 2024, paras. 70–77). EDPB guidance similarly treats factual decision-making power as more important than contractual labels (EDPB, 2021, paras. 12, 20, 28–30, 52–58).

Applied to the six-node model, the originating provider will ordinarily control generation and any service-side memory, retrieval or logging operations for which it determines the purposes and means. An enterprise integrator may be a joint controller where it jointly determines those matters, while a downstream institution that imports the allegation, creates its own risk field and makes a separate decision will generally be a controller for those operations (GDPR, 2016, art. 4(7); Fashion ID, 2019, paras. 65–85). Article 19 may require the relevant originating controller to communicate rectification to recipients, unless this proves impossible or involves disproportionate effort (GDPR, 2016, art. 19). Such

notification does not itself extend controllership to independent downstream processing; responsibility remains operation-specific (Fashion ID, 2019, paras. 68–85; IAB Europe, 2024, paras. 57–80). Responsibility is therefore allocated by processing operation and overlaps only where applicable law recognises overlap. Where processing is carried out by a processor, the rectification right is ordinarily exercised against the controller; Article 28(3)(e) requires processor assistance, and a processor that independently determines purposes and means is treated as controller for that processing (GDPR, 2016, arts. 16, 28(3)(e), 28(10)).

4.3 Why a single fix is inadequate

Hallucinated criminal claims may arise from unreliable training information, model-representation errors or incentives favouring guessing over abstention (Kalai et al., 2026). Retrieval may improve factuality but can also introduce errors through misleading context (Mallen et al., 2023). Model editing remains imperfect: corrections may fail across paraphrases, affect unrelated knowledge or degrade with repeated edits (Meng et al., 2022; Meng et al., 2023). MQuAKE further demonstrates that successful factual editing does not guarantee correct multi-hop reasoning (Zhong et al., 2023). Technical evaluations therefore assess reliability, generalisation, locality and stability rather than success on a single prompt (Pan et al., 2024; Wang et al., 2024).

Machine unlearning differs from rectification because it seeks removal of selected training influences, whereas rectification may concern runtime memory, retrieval documents or recipient databases (Cao and Yang, 2015; Bourtole et al., 2021). The framework therefore does not assume that parameter-level editing is always required. Technical intervention remains one possible means where evidence shows that model-level associations contribute materially to recurrence. The legal obligation remains tied to the relevant processing operation and applicable law.

4.4 The claimant's evidentiary position

The claimant may face structural evidentiary asymmetry because relevant system records and lifecycle documentation may be inaccessible to external reviewers, while organisations responsible for the system may possess internal models, data and process information unavailable externally (Cobbe et al., 2021; Raji et al., 2020). Contestable-design research supports accessible challenge routes (Alfrink et al., 2023), while audit research supports structured lifecycle documentation and review across scoping, mapping, artefact collection, testing and reflection (Raji et al., 2020). A credible notice should identify the affected person, the disputed meaning-bearing proposition, the relevant service or workflow, and *prima facie* evidence of falsity reasonably available to the claimant. Receipt of such notice should trigger preservation of relevant evidence and a proportionate internal inquiry.

Model cards, datasheets and AI factsheets can document system characteristics, data practices and governance arrangements (Mitchell et al., 2019; Gebru et al., 2021; Arnold et al., 2019). However, hidden technical dependencies may complicate error tracing (Sculley et al., 2015). Informational opacity should not strengthen a controller's evidential position where proportionate documentation could reasonably have reduced it. Accordingly, credible notice should trigger a practical evidential allocation rather than a

formal reversal of the legal burden: the claimant establishes *prima facie* falsity and a plausible service connection, while the actor controlling non-public evidence explains the nodes examined, interventions performed and justification for closure.

The framework therefore distinguishes the evidential threshold for opening a remedial inquiry from the threshold for propagation-wide rectification. *Prima facie* evidence of falsity is sufficient to trigger preservation, investigation and, where the conditions of applicable law are satisfied, interim restriction; it is not itself a final determination that the allegation is false. Where accuracy remains genuinely contested, Article 18(1)(a) GDPR provides for restriction of processing for a period enabling the controller to verify accuracy. Propagation-wide corrective measures should ordinarily follow a reasoned determination of inaccuracy based on the best authoritative evidence reasonably available (GDPR, 2016, art. 18(1)(a)).

5. The propagation-aware rectification standard

The framework can be stated in three propositions that organise the analysis that follows.

Proposition 1

Responsibility. Remedial responsibility attaches to processing operations for which an actor bears legally operative responsibility under the applicable regime, including operations whose purposes and means it determines alone or jointly; mere participation somewhere in the technical chain is insufficient (GDPR, 2016, art. 4(7); Fashion ID, 2019, paras. 68–85; IAB Europe, 2024, paras. 54–59, 70–77). Relevant factual influence may assist in determining such responsibility where applicable law adopts a functional rather than purely contractual test.

Proposition 2

Remedial reach. For each responsible actor, corrective action extends to reasonably identifiable nodes within its responsibility that retain a material capacity to continue, reproduce or transmit the verified false proposition, subject to proportionality and applicable legal limits.

Proposition 3

Closure. Under the proposed framework, rectification may be treated as legally complete when the responsible actor can provide reasonable and demonstrable assurance that required node-specific corrections and applicable recipient or decision review have been completed, proportionate recurrence testing has been conducted, and residual risks and limitations have been explained.

5.1 Three levels of correction

Output correction suppresses or changes a displayed answer. Proposition-level rectification targets materially equivalent assertions of the false factual predicate across aliases and paraphrases while

preserving accurate reporting, criticism and references to the original fabrication. For example, asserting that ‘Person A was convicted of Offence B’ differs from denying or critically discussing that claim. By analogy, the CJEU’s distinction between identical and equivalent unlawful content supports attention to semantic equivalence without requiring blanket suppression (Glawischnig-Piesczek, 2019, paras. 37–47). Testing should therefore cover relevant paraphrases and plausible multi-step pathways, as successful correction of one instance does not guarantee generalisation or correct reasoning across related inputs (Wang et al., 2024; Zhong et al., 2023).

Propagation-aware rectification adds the distributed dimension. Under the proposed framework, it requires proportionate inquiry into reasonably identifiable nodes within the actor’s responsibility, fulfilment of applicable recipient-notification duties, correction of inaccurate records, reassessment of materially affected outcomes where applicable law permits or requires, and evidence sufficient to justify closure. It is principally triggered in verified, high-severity cases where a credible propagation or reliance pathway exists and the sensitivity, foreseeable downstream use and potential material consequences justify the additional burden.

The standard is therefore deliberately asymmetric. A transient chatbot error that is promptly corrected, is not retained in persistent memory or retrieval, is not disclosed to an identifiable recipient, and has no plausible consequential use would ordinarily require output-level correction and proportionate documentation rather than a full propagation inquiry. The enhanced framework is triggered by verified severity plus a credible propagation or reliance pathway, not by factual error alone.

5.2 Correction provenance and a bounded stopping rule

A correction-provenance record links the credible notice, verified proposition, nodes examined, responsible actors, interventions performed, recipient communications, decision reviews, test results and unresolved residual risk. Data-provenance scholarship supplies concepts of origin, lineage and derivation (Buneman et al., 2001; Cheney et al., 2009), while PROV-DM provides a domain-agnostic vocabulary for entities, activities, derivations and responsible agents (Moreau and Missier, 2013); these concepts can be adapted to this remedial purpose. The record is not asserted to be an express freestanding GDPR right. It is a proposed evidentiary safeguard through which a controller can demonstrate compliance with existing accuracy, accountability and remedy duties. Its design must itself respect data minimisation, security and storage limitation (GDPR, 2016, arts. 5(1)(c)–(f), 24, 25): it should document remediation without becoming a searchable criminal dossier about the claimant. Access should therefore be tiered, with an intelligible summary available to the individual and confidential technical material available to a regulator or court where necessary (Dun & Bradstreet Austria, 2025, paras. 72–76).

Tracing relies on proportionately obtainable evidence, including logs, memory records, API transactions, recipient registers and decision files. Article 19’s impossibility and disproportionate-effort exception provides an analogy for the outer limit, while GDPR accuracy, accountability and design duties inform the proposed stopping rule (GDPR, 2016, arts. 5(1)(d), 19, 24, 25 and recitals 39, 66). A node is reasonably

identifiable where a documented or technically plausible pathway exists, the actor controls the relevant operation or can obtain information through enforceable arrangements, evidence is proportionately obtainable, and the node retains material capacity to harm. Tracing may cease where further inquiry is objectively impossible or disproportionate considering severity, continuing risk and expected remedial value. The actor must document the steps attempted, reasons for stopping and residual risks. This is an interpretive synthesis, not a claim that Article 19 creates a general propagation-tracing obligation.

Table 1 defines the principal thresholds used to trigger, bound and close propagation-aware rectification. References in this framework to an actor “bearing the burden” concern a proposed evidential burden of production when that actor invokes remedial closure; they do not purport to alter the formal burden of proof prescribed by applicable procedural law.

Table 1
Thresholds for Propagation-Aware Rectification

Concept	Working definition and principal factors	Burden and brief example
Credible notice	A request identifying the person, the disputed meaning-bearing proposition, the relevant service or workflow, and a <i>prima facie</i> basis for inaccuracy using evidence reasonably available to the claimant.	The claimant bears the initial burden of establishing a <i>prima facie</i> basis for falsity using evidence reasonably available to them. An official record contradicting an invented conviction will ordinarily satisfy that threshold, while a bare assertion may not.
Plausible propagation pathway	A technically or organisationally credible route by which the proposition could have moved from a known output into memory, retrieval, disclosure, a record, API output, or a decision.	The claimant shows a credible link; the actor controlling the relevant non-public processing evidence should explain the pathway within its responsibility. A matching risk field created after the relevant output may be sufficient to trigger inquiry.
Reasonably identifiable node	A node locatable through proportionately obtainable logs, versions, transaction records, data-flow documentation, recipient information, processor assistance, or workflow knowledge.	Under the proposed framework, the actor seeking closure should identify the nodes examined and explain the basis for excluding other plausible nodes from further inquiry. A known enterprise recipient is ordinarily identifiable; an unknown internet reader ordinarily is not.
Material reliance	Use of the false proposition with a realistic capacity to influence reasoning, procedural treatment, recommendation, or outcome; certainty of a different result is unnecessary.	Under the proposed framework, once a credible reliance link is shown, the actor responsible for the relevant decision process should investigate the extent of reliance. A criminal-risk flag used in screening is material; an unread attachment may not be.
Disproportionate effort	Tracing or communication whose cost, intrusion, or technical burden is excessive relative to severity, continuing risk, likelihood of finding a useful node, and available alternatives.	Under the proposed framework, the actor invoking disproportionate effort should document the methods considered, the steps attempted, the reasons further inquiry would be excessive, and any less burdensome alternatives. High-severity employment screening ordinarily justifies greater remedial effort than an ephemeral low-risk chat error.

Concept	Working definition and principal factors	Burden and brief example
Material recurrence	Reproduction or reintroduction of a materially equivalent false proposition during ordinary, foreseeable use after correction, through paraphrase, another node or a later version.	Under the proposed framework, the actor relying on closure should conduct recurrence testing proportionate to the identified risk and the technical pathways reasonably implicated by the case. Reappearance across routine fresh sessions is stronger evidence of material recurrence than a single remote adversarial prompt.
Reasonable and demonstrable assurance	Documented evidence that relevant nodes within the actor's responsibility were corrected, applicable recipient-communication duties were fulfilled, materially affected decisions were investigated, proportionate recurrence testing was completed, and residual risks and limitations were explained.	Under the proposed framework, the actor seeking closure bears the burden of demonstrating that the conditions for reasonable and demonstrable assurance have been satisfied. Assurance does not require proof of deletion from every model parameter or a guarantee against recurrence in every future system.

Source

Authors' synthesis from GDPR Articles 5, 16, 19, 24 and 25; relevant CJEU case law; EDPB guidance; and the interdisciplinary technical literature cited in Sections 4.3–4.4.

5.3 Recipients, decisions and reparation

Article 19 requires the controller to communicate rectification, erasure or restriction to each recipient to whom the personal data have been disclosed, unless this proves impossible or involves disproportionate effort (GDPR, 2016, art. 19). Where personal data have been or will be disclosed, Article 15(1)(c), as interpreted in *RW v Österreichische Post*, generally entitles the data subject to the actual identity of the recipients, unless identification is impossible or the access request is manifestly unfounded or excessive (*RW v Österreichische Post*, 2023, paras. 43, 46, 48–51). Under the proposed framework, such communication should identify the proposition, state that it was verified as false, request appropriate correction or restriction, and preserve evidence of the communication and any acknowledgement received. The originating provider is not required to control the recipient's independent system, but it must perform any applicable communication duty and, where required by law, cooperate with supervisory authorities.

Correcting data does not automatically repair an adverse decision already made. A material-reliance review should ask whether the proposition entered the decision, what weight it received, whether a human reviewer had real authority and access to contrary evidence, and whether the outcome could realistically have differed. Counterfactual analysis may help identify influential data and show how

changes could alter an outcome (Wachter et al., 2018). Whether a final decision must be reopened depends on the applicable legal regime; counterfactual analysis does not itself establish such a duty. Research on human oversight of government algorithms warns that nominal or ‘rubber-stamp’ human involvement may provide only superficial protection rather than meaningful review (Green, 2022), while work on human judgment in algorithmic loops distinguishes genuine discretion from formal sign-off (Binns, 2022). Where applicable law permits or requires, the institution should recalculate an intermediate score, reconsider a recommendation or reopen the final decision. If reopening is unavailable, the institution should pursue available lawful alternatives proportionate to the harm, which may include prospective correction, explanation, independent review and compensation where legally available.

Compensation addresses completed harm but does not replace cessation. CJEU case law requires proof of damage and causation but rejects an autonomous seriousness threshold (*Österreichische Post*, 2023, paras. 32, 42, 50–51), recognises that a well-founded fear of future misuse may constitute non-material damage (*Natsionalna agentsia za prihodite*, 2023, paras. 79–86), and permits minimal compensation where the proven damage is not serious, provided that the award fully compensates the damage suffered (*Scalable Capital*, 2024, paras. 23–24, 43–46). An effective remedial package may therefore combine correction, decision review, declaratory relief and monetary reparation.

Table 2 calibrates the required inquiry and remedial response according to the degree of material reliance on the false proposition.

Table 2
Material Reliance and Proportionate Remedial Responses

Reliance situation	Required inquiry	Proportionate remedial response
False data was present but not used	Verify that the proposition did not enter reasoning, recommendation, procedural treatment, or outcome.	Correct or restrict the data and document non-reliance; renewed review of the decision is ordinarily unnecessary.
False data affected procedure but not the final outcome	Determine whether it changed scrutiny, delay, burden, explanation, or the opportunity to participate.	Correct the record, repair the affected procedure where feasible, and provide a revised explanation.
False data materially affected a recommendation or intermediate score	Assess assigned weight, causal proximity, and downstream reliance.	Recalculate or reassess the recommendation and, where an applicable recipient-communication duty attaches, communicate the correction to the relevant recipient.
False data materially influenced a final decision	Determine whether the result or treatment could realistically have differed without the false proposition.	Reopen, repeat, or reconsider where applicable law permits or requires; preserve access to supervisory or judicial review where such review is available.
A human allegedly verified the automated result	Examine the reviewer's authority and competence, access to contrary evidence, available time and freedom to depart from the automated result.	Where the purported human review was merely nominal, the proposed framework treats it as insufficient to establish meaningful independent scrutiny; a fresh reassessment should be undertaken where applicable law or the relevant remedial process permits or requires.
Renewed review of the decision is legally or practically unavailable	Identify why renewed review is unavailable and whether another measure can mitigate continuing or completed harm.	Use available lawful alternatives proportionate to the harm, which may include a correction notice, prospective exclusion, independent review, explanation, compensation where legally available, or another appropriate remedy.

Source: Authors' synthesis from GDPR Articles 16, 19, 22 and 82; SCHUFA; Dun & Bradstreet Austria (2025); compensation jurisprudence; and peer-reviewed scholarship on meaningful human oversight.

5.4 Recurrence testing and reasoned closure

Recurrence may be semantic, where a paraphrase reproduces the false proposition, or cascading, where an uncorrected node reintroduces it. Proportionate testing should cover relevant name variants, fresh sessions, memory, retrieval, tool calls and material system updates where technically applicable. Interventions must also preserve accurate information about genuine proceedings, acquittals and public controversy. Model-editing research demonstrates the importance of paraphrase, locality and multi-hop validation (Meng et al., 2022; Zhong et al., 2023). Retrieval requires separate testing where it presents a plausible recurrence pathway (Mallen et al., 2023). Reappearance may justify further inquiry but does not

independently identify its cause. Where feasible, provenance records, retrieval traces and logs should support attribution of recurrence to particular processing nodes.

Table 3 consolidates the five-stage post-notice rectification test and distinguishes express legal duties, interpretive applications and proposed evidentiary safeguards.

Table 3
Five-Stage Post-Notice Rectification Test

Stage	Required inquiry and action	Human-rights function, record and legal character
Notice and containment	Identify the person, proposition, workflow, <i>prima facie</i> contradiction and continuing risk; preserve evidence and use a narrowly tailored interim restriction where justified.	Access and cessation. Express legal duties, supplemented by interpretive interim containment.
Determination and mapping	Determine accuracy and identify reasonably identifiable nodes at which the proposition was stored, transmitted, transformed or materially used.	Reasoned finding and bounded map. Interpretive application of accuracy, rectification, accountability and design duties.
Allocation and correction	Assign each node to the responsible actor and implement node-specific correction, restriction, or processor assistance.	Actor-node allocation and action record. Express duties are allocated by operation; coordination safeguards are partly proposed.
Recipients and decisions	Fulfil applicable recipient-communication duties and investigate whether the false data materially affected an assessment or outcome.	Article 19 communication and reliance assessment. Decision consequences depend on Article 22 and applicable law.
Testing and closure	Test semantic, cascading and temporal recurrence; explain remedial reach, limits, residual risk and the route to renewed review where available under applicable law.	Implementation and non-recurrence. Correction provenance and dated tests are proposed evidentiary safeguards.

Source: Authors' synthesis of the five remedial functions outlined in Section 3.2, together with the international, regional and domestic legal authorities cited across Sections 3–5.

A reasoned closure decision should contain a minimum evidentiary package: the verified inaccuracy; a bounded map of nodes examined; an action log recording correction or restriction; applicable recipient-notification records; any material-reliance or decision review; dated recurrence tests identifying the system versions covered; and a residual-risk statement with the route to renewed review where available under applicable law. This package is not proposed as a freestanding right. It is the evidence by which an actor invoking closure can demonstrate that the relevant legal duties and proposed safeguards were actually implemented. Closure should not be treated as adequately demonstrated where a material, reasonably identifiable node remains unexamined without sufficient justification, and it should remain subject, where available, to independent supervisory or judicial review.

Remedial closure is therefore time- and version-bounded rather than a permanent certification of the system. A material change to the model, retrieval source, memory architecture or downstream workflow, or credible evidence of renewed material recurrence, may justify reopening the inquiry even though closure was reasonably established on the earlier evidential record.

6. Application, proportionality and limits

6.1 Applying the standard to the Holmen scenario

The Holmen complaint provides the factual starting point (Noyb, 2025), while the six-node chain remains an illustrative scenario rather than a reconstruction of the provider’s architecture. The disputed proposition alleged that Holmen had been convicted of killing two children, attempting to kill a third and receiving a twenty-one-year sentence. Under the proposed framework, credible evidence of falsity should trigger evidence preservation and a reasoned accuracy determination. If fabrication is confirmed and a credible memory or retrieval pathway could regenerate the allegation, output blocking alone is insufficient. Technical research supports this concern because factual editing may fail across multi-hop reasoning, while erroneous agent-memory records can influence later outputs (Zhong et al., 2023; Xiong et al., 2026).

The originating provider’s inquiry should cover the generated output and, where technically relevant and reasonably identifiable, service-side memory, retrieval components, relevant logs and identifiable API disclosures within its responsibility. For each such node, it should record whether the proposition was present, the corrective action, verification method and resulting status. If enterprise recipients received the proposition, Article 19 requires the relevant controller to communicate rectification unless this proves impossible or involves disproportionate effort (GDPR, 2016, art. 19). A recipient that creates its own criminal-risk field or separate decision may independently be controller for that processing and bear the duties applicable to it. Where relevant technical information contains protected third-party data or trade secrets, that information may require disclosure to the competent supervisory authority or court for independent balancing and review (Dun & Bradstreet Austria, 2025, paras. 67–76). Unresolved technical limitations should be documented where appropriate (NIST, 2023). Table 4 translates the six-node scenario into an illustrative correction-provenance record for each responsible operation.

Applied to the Holmen scenario, the framework would not support remedial closure where a reasonably identifiable and materially relevant processing node remains unexamined without adequate justification. Closure could instead be established through documented, proportionate corrective measures within each responsible actor’s legal and operational authority, together with appropriate verification and an explanation of residual risks. This conclusion does not require proof that recurrence is impossible.

Table 4
Illustrative Six-Node Correction-Provenance Record

Node and responsible operation	Node-specific response	Evidence and closure status
Generated output: originating provider	Correct or suppress the reported answer and preserve the incident record.	Corrected output, incident identifier and dated implementation log.
Service-side memory: originating provider	Invalidate or amend the memory item and prevent routine retrieval of the fabricated proposition.	Memory event record and fresh-session tests; status: corrected or residual risk explained.
Retrieval index: actor responsible for the retrieval or indexing operation	Correct, qualify or exclude the source or indexed entry responsible for the false status.	Index update, retrieval trace and tests using ordinary variants of the person's name.
API disclosure: originating provider; known recipient assessed separately for its own processing	Communicate rectification under applicable recipient-notification duty and preserve delivery evidence.	Recipient identity, notice date and a status such as sent, received, matched or failed.
Downstream risk field: downstream controller	Remove or qualify the criminal-risk field and identify uses of the field in screening or assessment.	Database audit, corrected field and documented reliance investigation.
Consequential decision: decision controller	Reassess or reopen where applicable law permits or requires; otherwise pursue available lawful alternatives proportionate to the harm where the false proposition materially influenced treatment.	Revised decision or a reasoned determination that renewed review is unavailable, with access to supervisory or judicial review preserved where available under applicable law.

Source

Authors' application of the propagation-aware rectification framework to the six-node scenario, informed by the factual parameters of the March 2025 Noyb complaint against OpenAI.

6.2 Institutional pathway and independent verification

An effective institutional pathway begins with an accessible complaint channel and competent human review. The UNGPs and AI Convention support grievance procedures capable of practical relief (OHCHR, 2011, principles 29, 31; Council of Europe, 2024, arts. 14–15). Internal audits can support accountability and evidence preservation but do not replace independent review where required (Raji et al., 2020; Wieringa, 2020).

Regulators should have graduated powers: urgent restriction and evidence-preservation orders for immediate high-risk harm; binding rectification, recipient-notification and audit orders after investigation; and proportionate penalties for obstruction or repeated non-compliance. Courts should be able, within

the powers available under applicable law, to examine confidential architecture evidence, order correction or reassessment, award compensation where appropriate, and preserve access to renewed review if the false allegation reappears.

6.3 Freedom of expression and abuse prevention

Where ECHR Article 10 is engaged, propagation-aware rectification must respect freedom of expression. The framework targets verified false criminal-status allegations, not opinion, satire, public-interest reporting or accurate criminal history. *Thomaidis v Greece* confirms that Article 10 does not immunise irresponsible dissemination of confidential investigative material where domestic courts conduct a Convention-compliant balance (*Thomaidis v Greece*, 2024, paras. 27, 30, 35–38). Węgrzynowski cautions against rewriting history through wholesale removal of archived publications and recognises contextual qualification, such as adding a reference to later judgments, as a potentially less restrictive response (*Węgrzynowski and Smolczewski v Poland*, 2013, paras. 59, 65–68).

Procedural safeguards should include proportionate identity verification, identification of the disputed proposition, evidence of falsity, preservation of accurate public-interest material and measures against demonstrably fraudulent requests. GDPR Article 10 raises a separate doctrinal question. *GC and Others v CNIL* treats search results linked to reports of criminal proceedings as data relating to offences and convictions, and *Darashev* confirms that data recording a person’s status as a suspect may fall within Article 10 even where commission of the suspected offence was not established (*GC and Others v CNIL*, 2019, paras. 72–73; *Darashev*, 2026, para. 106). Whether an entirely invented conviction falls within Article 10 remains open, but that uncertainty does not remove the general duties of accuracy and rectification (GDPR, 2016, arts. 5(1)(d), 16).

6.4 Limitations

The framework has four principal limits. First, proprietary architectures differ, so the six-node model cannot represent every deployment. Secondly, the analysis lacks confidential logs and unredacted complaint files and does not establish what occurred inside any particular provider. Thirdly, model-correction and unlearning evaluation remains unsettled: automated-correction research still lacks comprehensive quantitative evaluation and may show limited generalisation, while unlearning can fail under indirect or rephrased queries (Pan et al., 2024; Maini et al., 2024; Thaker et al., 2025). Fourthly, the scenario cannot quantify compliance costs or cross-border enforcement burdens. The study therefore tests the internal legal and technical coherence of a doctrinally grounded normative standard; it does not claim empirical validation of operational effectiveness, comparative cost or scalability.

The proposal is also deliberately narrow. It applies most strongly to verified, high-severity fabricated criminal allegations where the information is sensitive, a credible propagation or reliance pathway exists, downstream reliance is foreseeable, and the evidence of falsity is strong. It does not impose a zero-error duty for every trivial hallucination. Further research should test the framework against regulatory case files, actual data-flow documentation and recurrence results across system versions and provider sizes.

7. Discussion

7.1 Remedial closure as the missing legal object

The framework provides a bounded test of remedial closure integrating propagation, actor-specific responsibility, node-specific intervention, evidentiary verification and a stopping rule. Its doctrinal and conceptual contribution is to define the endpoint of remediation, allocate evidential production under information asymmetry and make closure inspectable by supervisory or judicial bodies where applicable law provides review.

The framework also distinguishes three forms of completeness. Doctrinal completeness means that the responsible actor has satisfied the legal duties applicable to its processing operations. Evidentiary completeness means that it can demonstrate, through a bounded node map, action records and proportionate testing, why closure is justified. Absolute technical completeness would require elimination of the false proposition from every conceivable system and future pathway. Remedial closure requires the first two, supported by proportionate technical validation within the actor's authority; it does not require absolute technical completeness. This distinction allows documented residual risk to coexist with closure without reducing closure to a provider's unsupported assertion, and it explains why the framework rejects both output-only correction and strict chain-wide liability.

Although developed for fabricated criminal allegations, the framework may inform analysis of other verified, high-severity inaccuracies, such as false professional disqualifications or fabricated financial misconduct. Transfer is not automatic: legal basis, data sensitivity, suppression risks and remedial intensity must be reassessed in each domain. The claim is functional generalisability of the analytical questions, not automatic portability of resulting obligations.

7.2 Cross-border propagation and jurisdictional limits

Cross-border propagation exposes the limits of any single legal order. A false proposition may move from one provider through an API to an enterprise in another jurisdiction and then into a separate decision system. The framework does not assume universal territorial reach or chain-wide technical closure. Under EU law, Article 3 GDPR determines the Regulation's territorial scope, while Articles 44–49 govern transfers of personal data to third countries and international organisations; Article 19 may additionally require recipient communication without making the originating controller responsible for independent downstream processing (GDPR, 2016, arts. 3, 19, 44–49). The PIPL likewise contains specified extraterritorial rules (PIPL, 2021, art. 3). An originator may therefore reach doctrinal and evidentiary closure within its legally and operationally reachable domain even though residual risk persists at a foreign independent recipient. That conclusion requires completion of applicable notification duties, documentation of the inaccessible or independent node, and use of available supervisory or judicial cooperation where required; it is actor-level remedial closure, not a representation that every downstream manifestation has disappeared.

8. Conclusion

An AI-fabricated criminal allegation can become an infrastructural criminal identity when it persists beyond the visible output and enters records, risk classifications or consequential decisions. Output-only suppression may therefore be formally available yet practically ineffective. Human-rights law supplies principles of cessation and effective redress; business-and-human-rights standards support accessible grievance and cooperation in remediation; and data-protection law supplies operational duties concerning accuracy, access, rectification, recipient communication, accountability and, where legal conditions are met, compensation. The central problem is completion: when may a responsible actor, regulator or court justifiably conclude that the remedy has worked?

Propagation-aware rectification supplies a bounded answer: accessible challenge and reasoned determination, cessation at reasonably identifiable nodes, correction of inaccurate records, reassessment of materially affected decisions where applicable law permits or requires, verifiable implementation and proportionate non-recurrence testing. Responsibility is assigned by processing operation. The claimant identifies the disputed proposition, establishes *prima facie* falsity and shows a plausible service connection; the actor controlling relevant non-public evidence explains the scope and outcome of its inquiry.

The standard rejects universal erasure from every possible model, indefinite tracing of hypothetical dissemination and strict liability for independent downstream actors. Closure instead requires reasonable, demonstrable assurance that each responsible actor has corrected or contained the verified false allegation across reasonably identifiable, materially relevant nodes within its authority, addressed applicable downstream effects and documented residual risks. Where that conclusion is supported by an inspectable correction-provenance record and remains subject, where available, to independent supervisory or judicial review, rectification may be treated as legally complete.

Declarations

Competing interests

The author declares no competing interests.

Ethics declaration

Ethical approval

was not required as this research did not involve human participants or human data.

Author Contribution

H.M. contributed to Conceptualization, Methodology, Formal Analysis, Investigation, Writing – Original Draft, Writing – Review & Editing, and Project Administration. X.G. contributed to Conceptualization, Methodology, Supervision, Validation, Project Administration, and Writing – Review & Editing. Y.L. contributed to Formal Analysis, Investigation, Resources, and Writing – Review & Editing. All authors reviewed and approved the manuscript.

Acknowledgement

The authors gratefully acknowledge the Institute for Human Rights at Southeast University, China, and thank its members for the valuable support and discussions that contributed to the preparation of this manuscript.

Data Availability

Data sharing is not applicable to this research as no data were generated or analysed.

References

1. African Charter on Human and Peoples' Rights (1981) Adopted 27 June 1981, entered into force 21 October 1986, 1520 UNTS 217
2. Alfrink K, Keller I, Kortuem G et al (2023) Contestable AI by design: towards a framework. *Minds Mach* 33:613–639. <https://doi.org/10.1007/s11023-022-09611-z>
3. American Convention on Human Rights (1969) Adopted 22 November 1969, entered into force 18 July 1978, 1144 UNTS 123
4. Arnold M, Bellamy RKE, Hind M et al (2019) FactSheets: increasing trust in AI services through supplier's declarations of conformity. *IBM J Res Dev* 63(4/5):6:1–613. <https://doi.org/10.1147/JRD.2019.2942288>
5. Article 29 Data Protection Working Party (2018) Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679. WP251rev.01, European Commission, revised and adopted 6 February 2018
6. Beduschi A (2026) Data protection in the era of agentic artificial intelligence. *Comput Law Secur Rev* 61:106342. <https://doi.org/10.1016/j.clsr.2026.106342>
7. Bender EM, Gebru T, McMillan-Major A et al (2021) On the dangers of stochastic parrots: can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, pp 610–623. <https://doi.org/10.1145/3442188.3445922>
8. *Biancardi v Italy* (2021) ECtHR App No 77419/16, judgment 25 November 2021

9. Binns R (2022) Human judgment in algorithmic loops: individual justice and automated decision-making. *Regul Gov* 16:197–211. <https://doi.org/10.1111/rego.12358>
10. Borgeaud S, Mensch A, Hoffmann J et al (2022) Improving language models by retrieving from trillions of tokens. In: *Proceedings of the 39th International Conference on Machine Learning*. *Proceedings of Machine Learning Research*, vol 162. PMLR, pp 2206–2240
11. Bourtole L, Chandrasekaran V, Choquette-Choo CA et al (2021) Machine unlearning. In: *2021 IEEE Symposium on Security and Privacy*. IEEE, pp 141–159. <https://doi.org/10.1109/SP40001.2021.00019>
12. Buneman P, Khanna S, Tan W-C (2001) Why and where: a characterization of data provenance. In: Van den Bussche J, Vianu V (eds) *Database Theory—ICDT 2001*. *Lecture Notes in Computer Science*, vol 1973. Springer, Heidelberg, pp 316–330. https://doi.org/10.1007/3-540-44503-X_20
13. Cao Y, Yang J (2015) Towards making systems forget with machine unlearning. In: *2015 IEEE Symposium on Security and Privacy*. IEEE, pp 463–480. <https://doi.org/10.1109/SP.2015.35>
14. Cheney J, Chiticariu L, Tan W-C (2009) Provenance in databases: why, how, and where. *Found Trends Databases* 1(4):379–474. <https://doi.org/10.1561/19000000006>
15. Citron DK (2008) Technological due process. *Wash Univ Law Rev* 85:1249–1313
16. Citron DK, Solove DJ (2022) Privacy harms. *Boston Univ Law Rev* 102:793–863
17. Cobbe J, Lee MSA, Singh J (2021) Reviewable automated decision-making: a framework for accountable algorithmic systems. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, pp 598–609. <https://doi.org/10.1145/3442188.3445921>
18. Council of Europe (2024) Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law. CETS No 225, opened for signature in Vilnius on 5 September 2024
19. Council of Europe Committee of Ministers (2020) Recommendation CM/Rec(2020)1 on the human rights impacts of algorithmic systems, adopted 8 April 2020
20. Crawford K, Schultz J (2014) Big data and due process: toward a framework to redress predictive privacy harms. *Boston Coll Law Rev* 55:93–128
21. CRIF (2023) Case C-487/21, Österreichische Datenschutzbehörde and CRIF, EU:C:2023:369, judgment 4 May 2023
22. Cyberspace Administration of China (2025) Measures for the Administration of Personal Information Protection Compliance Audits. Order No 18, promulgated 12 February 2025, effective 1 May 2025.
23. Cyberspace Administration of China, Ministry of Industry and Information Technology, Ministry of Public Security, State Administration for Market Regulation (2021) Provisions on the Administration of Algorithmic Recommendation in Internet Information Services. Order 9, promulgated 31 December 2021, effective 1 March 2022.

24. Cyberspace Administration of China, Ministry of Industry and Information Technology, Ministry of Public Security (2022) Provisions on the Administration of Deep Synthesis Internet Information Services. Order 12, promulgated 25 November 2022, effective 10 January 2023.
25. Darashev (2026) Case C-312/24, CL v Prokuratura na Republika Bulgaria, EU:C:2026:449, judgment 4 June 2026
26. Deldits (2025) Case C-247/23, VP v Országos Idegenrendészeti Főigazgatóság, EU:C:2025:172, judgment 13 March 2025
27. Digital Omnibus on AI (2026) Regulation (EU) 2026/1744 of the European Parliament and of the Council of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI). OJ L, 2026/1744, 24.7.2026
28. Dun & Bradstreet Austria (2025) Case C-203/22, EU:C:2025:117, judgment 27 February 2025
29. ECHR (1950) Convention for the Protection of Human Rights and Fundamental Freedoms. ETS No 5, Rome, 4 November 1950
30. EDPB (2020) Guidelines 4/2019 on Article 25 data protection by design and by default, version 2.0, adopted 20 October 2020. https://www.edpb.europa.eu/documents/guideline/guidelines-42019-on-article-25-data-protection-by-design-and-by-default_en. Accessed 29 Aug 2026
31. EDPB (2021) Guidelines 07/2020 on the concepts of controller and processor in the GDPR, version 2.1, adopted 7 July 2021, minor corrections 20 September 2022. https://www.edpb.europa.eu/documents/guideline/guidelines-072020-on-the-concepts-of-controller-and-processor-in-the-gdpr_en. Accessed 29 Aug 2026
32. EDPB (2023) Guidelines 01/2022 on data subject rights – right of access, version 2.1, adopted 28 March 2023. https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-012022-data-subject-rights-right-access_en. Accessed 29 Aug 2026
33. EDPB (2024a) Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, adopted 17 December 2024. https://www.edpb.europa.eu/documents/opinion-of-the-board-art-64/opinion-282024-on-certain-data-protection-aspects-related-to_en. Accessed 29 Aug 2026
34. EDPB (2024b) Report of the work undertaken by the ChatGPT Taskforce, adopted 23 May 2024. https://www.edpb.europa.eu/system/files/2024-05/edpb_20240523_report_chatgpt_taskforce_en.pdf. Accessed 29 Aug 2026
35. Edwards L, Veale M (2017) Slave to the algorithm? Why a ‘right to an explanation’ is probably not the remedy you are looking for. *Duke Law Technol Rev* 16:18–84
36. EU AI Act (2024) Regulation (EU) 2024/1689 (Artificial Intelligence Act). OJ L, 2024/1689, 12.7.2024
37. EU Charter (2016) Charter of Fundamental Rights of the European Union. OJ C 202, 7.6.2016, pp 389–405
38. Fashion ID (2019) Case C-40/17, Fashion ID GmbH & Co. KG v Verbraucherzentrale NRW eV, EU:C:2019:629, judgment 29 July 2019

39. GC and Others v CNIL (2019) Case C-136/17, GC, AF, BH and ED v Commission nationale de l'informatique et des libertés (CNIL), EU:C:2019:773, judgment 24 September 2019
40. GDPR (2016) Regulation (EU) 2016/679 (General Data Protection Regulation). OJ L 119, 4.5.2016, pp 1–88
41. Gebru T, Morgenstern J, Vecchione B et al (2021) Datasheets for datasets. *Commun ACM* 64(12):86–92. <https://doi.org/10.1145/3458723>
42. Generative AI, Measures (2023) Shengchengshi rengong zhineng fuwu guanli zanxing banfa (生成式人工智能服务管理规定; Interim Measures for the Management of Generative Artificial Intelligence Services). Order No 15, promulgated 10 July 2023, effective 15 August 2023.
43. Glawischnig-Piesczek (2019) Case C-18/18, Eva Glawischnig-Piesczek v Facebook Ireland Limited, EU:C:2019:821, judgment 3 October 2019
44. Goldstein A (2025) Agentic AI. In: European Data Protection Supervisor, TechSonar Report 2025–2026. Publications Office of the European Union, pp 1–5. <https://doi.org/10.2804/8205191>
45. Google (Dereferencing of Allegedly Inaccurate Content) (2022) Case C-460/20, TU and RE v Google LLC, EU:C:2022:962, judgment 8 December 2022
46. Google S (2014) Case C-131/12, Google Spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González, EU:C:2014:317, judgment 13 May 2014
47. Graßhoff H, Hansen M, Jensen M et al (2026) Short paper: Models in the dark—rectification and erasure under GDPR in ML supply chains. In: Pallas F, Leschke N, Arastouei N, Rannenberg K (eds) Privacy technologies and policy. APF 2026. Lecture Notes in Computer Science, vol 16847. Springer, Cham, pp 71–85. https://doi.org/10.1007/978-3-032-35899-8_4
48. Green B (2022) The flaws of policies requiring human oversight of government algorithms. *Comput Law Secur Rev* 45:105681. <https://doi.org/10.1016/j.clsr.2022.105681>
49. Guu K, Lee K, Tung Z et al (2020) Retrieval augmented language model pre-training. In: Daumé H III, Singh A (eds) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol 119. PMLR, pp 3929–3938
50. Häuselmann A, Custers B (2024) The right to rectification and inferred personal data. *Eur J Law Technol* 15(3):1–24
51. Herrerias Castro L (2026) Legal remedies for AI-generated reputational harm. *Camb Forum AI: Law Gov* 2:e12. <https://doi.org/10.1017/cfl.2026.10062>
52. Human Rights Committee (2004) General Comment No 31: The nature of the general legal obligation imposed on States Parties to the Covenant. UN Doc CCPR/C/21/Rev.1/Add.13, 26 May 2004
53. Hurbain v Belgium (2023) [GC] ECtHR App 57292/16, judgment 4 July 2023.
54. IAB Europe (2024) Case C-604/22, IAB Europe v Gegevensbeschermingsautoriteit, EU:C:2024:214, judgment 7 March 2024

55. ICCPR (1966) International Covenant on Civil and Political Rights. Adopted 16 December 1966, entered into force 23 March 1976, 999 UNTS 171
56. Jehovan todistajat (2018) Case C-25/17, Tietosuoja-valtuutettu v Jehovan todistajat — uskonnollinen yhdyskunta, EU:C:2018:551, judgment 10 July 2018
57. Ji Z, Lee N, Frieske R et al (2023) Survey of hallucination in natural language generation. *ACM Comput Surv* 55(12):248:1–24838. <https://doi.org/10.1145/3571730>
58. Kalai AT, Nachum O, Vempala SS et al (2026) Evaluating large language models for accuracy incentivizes hallucinations. *Nature* 653:1047–1051. <https://doi.org/10.1038/s41586-026-10549-w>
59. Kaminski ME (2019) The right to explanation, explained. *Berkeley Technol Law J* 34:189–218. <https://doi.org/10.15779/Z38TD9N83H>
60. *Khelili v Switzerland* (2011) ECtHR App No 16188/07, judgment 18 October 2011
61. Kinchin N (2024) Voiceless’: the procedural gap in algorithmic justice. *Int J Law Inf Technol* 32:eaae024. <https://doi.org/10.1093/ijlit/eaae024>
62. Kroll JA, Huey J, Barocas S et al (2017) Accountable algorithms. *Univ Pa Law Rev* 165:633–705
63. Lewis P, Perez E, Piktus A et al (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv Neural Inf Process Syst* 33:9459–9474
64. Lin S, Hilton J, Evans O (2022) TruthfulQA: measuring how models mimic human falsehoods. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, pp 3214–3252. <https://doi.org/10.18653/v1/2022.acl-long.229>
65. *Germany WW* (2018) ECtHR App Nos 60798/10 and 65599/10, judgment 28 June 2018.
66. *M.M. v United Kingdom* (2012) ECtHR App No 24029/07, judgment 13 November 2012
67. Maini P, Feng Z, Schwarzschild A et al (2024) TOFU: a task of fictitious unlearning for LLMs. In: *First Conference on Language Modeling (COLM 2024)*. <https://openreview.net/forum?id=B41hNBowLo>. Accessed 28 Aug 2026
68. Mallen A, Asai A, Zhong V et al (2023) When not to trust language models: investigating effectiveness of parametric and non-parametric memories. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, pp 9802–9822. <https://doi.org/10.18653/v1/2023.acl-long.546>
69. Manakul P, Liusie A, Gales MJF (2023) SelfCheckGPT: zero-resource black-box hallucination detection for generative large language models. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, pp 9004–9017. <https://doi.org/10.18653/v1/2023.emnlp-main.557>
70. *Margari v Greece* (2023) ECtHR App No 36705/16, judgment 20 June 2023
71. McGregor L, Murray D, Ng V (2019) International human rights law as a framework for algorithmic accountability. *Int Comp Law Q* 68:309–343. <https://doi.org/10.1017/S0020589319000046>

72. Meng K, Bau D, Andonian A et al (2022) Locating and editing factual associations in GPT. *Adv Neural Inf Process Syst* 35:17359–17372
73. Meng K, Sharma AS, Andonian AJ et al (2023) Mass-editing memory in a transformer. In: *The Eleventh International Conference on Learning Representations (ICLR 2023)*. OpenReview.net. <https://openreview.net/forum?id=MkbcAHlYgyS>. Accessed 29 Aug 2026
74. Min S, Krishna K, Lyu X et al (2023) FActScore: fine-grained atomic evaluation of factual precision in long-form text generation. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, pp 12076–12100. <https://doi.org/10.18653/v1/2023.emnlp-main.741>
75. Mitchell M, Wu S, Zaldivar A et al (2019) Model cards for model reporting. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, New York, pp 220–229. <https://doi.org/10.1145/3287560.3287596>
76. Moreau L, Missier P (eds) (2013) PROV-DM: the PROV data model. W3C Recommendation, 30 April 2013. <https://www.w3.org/TR/prov-dm/>. Accessed 29 Aug 2026
77. Natsionalna agentsia za prihodite (2023) Case C-340/21, EU:C:2023:986, judgment 14 December 2023
78. NIST (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1, 26 January 2023. <https://doi.org/10.6028/NIST.AI.100-1>
79. NIST (2024) Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1, 26 July 2024. <https://doi.org/10.6028/NIST.AI.600-1>
80. Nowak (2017) Case C-434/16, EU:C:2017:994, judgment 20 December 2017
81. Noyb (2025) Complaint against OpenAI OpCo, LLC, Case No C096, 20 March 2025. https://noyb.eu/sites/default/files/2025-03/OpenAI_complaint_redacted.pdf. Accessed 28 Aug 2026
82. OECD (2024) Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449, adopted 22 May 2019, amended 8 November 2023 and 3 May 2024
83. Office of the Privacy Commissioner of Canada, Commission d'accès à l'information du Québec, Office of the Information and Privacy Commissioner for British Columbia et al et al (2026) Joint investigation of OpenAI OpCo, LLC. PIPEDA Report of Findings #2026-002, 6 May 2026. <https://www.priv.gc.ca/en/opc-actions-and-decisions/investigations/investigations-into-businesses/2026/pipeda-2026-002/>. Accessed 28 Aug 2026
84. OHCHR (2011) Guiding principles on business and human rights: implementing the United Nations 'Protect, Respect and Remedy' framework. United Nations, New York and Geneva. UN Doc HR/PUB/11/04
85. Österreichische Post (2023) Case C-300/21, EU:C:2023:370, judgment 4 May 2023
86. Pan L, Saxon M, Xu W et al (2024) Automatically correcting large language models: surveying the landscape of diverse automated correction strategies. *Trans Assoc Comput Linguist* 12:484–506. https://doi.org/10.1162/tacl_a_00660

87. Pankki S (2023) Case C-579/21, EU:C:2023:501, judgment 22 June 2023
88. Park JS, O'Brien JC, Cai CJ et al (2023) Generative agents: interactive simulacra of human behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. ACM, New York, art 2, pp 1–22. <https://doi.org/10.1145/3586183.3606763>
89. PIPL (2021) Personal Information Protection Law of the People's Republic of China. Adopted 20 August 2021, effective 1 November 2021
90. PRC Civil Code (2020) Civil Code of the People's Republic of China. Adopted 28 May 2020, effective 1 January 2021
91. Raji ID, Smart A, White RN et al (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. ACM, New York, pp 33–44. <https://doi.org/10.1145/3351095.3372873>
92. Regulations on Network Data Security Management (2024) State Council Decree No 790, adopted 30 August 2024, promulgated 24 September 2024, effective 1 January 2025
93. Rotaru v Romania (2000) [GC] ECtHR App 28341/95, judgment 4 May 2000.
94. RW v Österreichische Post (2023) Case C-154/21, EU:C:2023:3, judgment 12 January 2023
95. S and Marper v United Kingdom (2008) [GC] ECtHR App Nos 30562/04 and 30566/04, judgment 4 December 2008.
96. Scalable Capital (2024) Joined Cases C-182/22 and C-189/22, EU:C:2024:531, judgment 20 June 2024.
97. SCHUFA Holding (2023) Case C-634/21, EU:C:2023:957, judgment 7 December 2023
98. Sculley D, Holt G, Golovin D et al (2015) Hidden technical debt in machine learning systems. *Adv Neural Inf Process Syst* 28:2503–2511
99. Shinn N, Cassano F, Gopinath A et al (2023) Reflexion: language agents with verbal reinforcement learning. *Adv Neural Inf Process Syst* 36:8634–8652. <https://doi.org/10.52202/075280-0377>
100. Söderman v Sweden (2013) [GC] ECtHR App No 5786/08, judgment 12 November 2013
101. Supreme People's Court (2021) Provisions on Several Issues Concerning the Application of Law in Civil Cases Involving the Use of Facial Recognition Technologies to Process Personal Information, Judicial Interpretation 15, adopted 8 June 2021, effective 1 August 2021.
102. Thaker P, Hu S, Kale N et al (2025) Position: LLM unlearning benchmarks are weak measures of progress. In: 2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). IEEE, pp 520–533. <https://doi.org/10.1109/SaTML64287.2025.00035>
103. Thomaidis v Greece (2024) ECtHR App No 28345/16, judgment 7 May 2024
104. UN General Assembly (2005) Basic Principles and Guidelines on the Right to a Remedy and Reparation for Victims of Gross Violations of International Human Rights Law and Serious Violations of International Humanitarian Law. UNGA Res 60/147, UN Doc A/RES/60/147, 16 December 2005.

105. UNESCO (2021) Recommendation on the ethics of artificial intelligence. UNESCO, Paris. adopted 23 November 2021
106. Veale M, Zuiderveen Borgesius F (2021) Demystifying the draft EU Artificial Intelligence Act. *Comput Law Rev Int* 22:97–112. <https://doi.org/10.9785/cri-2021-220402>
107. Wachter S, Mittelstadt B, Russell C (2018) Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harv J Law Technol* 31:841–887
108. Wang P, Zhang N, Tian B et al (2024) EasyEdit: an easy-to-use knowledge editing framework for large language models. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Association for Computational Linguistics, Bangkok, pp 82–93. <https://doi.org/10.18653/v1/2024.acl-demos.9>
109. Węgrzynowski and Smolczewski v Poland (2013) ECtHR App No 33846/07, judgment 16 July 2013.
110. Weidinger L, Uesato J, Rauh M et al (2022) Taxonomy of risks posed by language models. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, pp 214–229. <https://doi.org/10.1145/3531146.3533088>
111. Wieringa M (2020) What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In: Hildebrandt M, Castillo C (eds) *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, New York, pp 1–18. <https://doi.org/10.1145/3351095.3372833>
112. Wirtschaftsakademie (2018) Case C-210/16, EU:C:2018:388, judgment 5 June 2018
113. Xiong Z, Lin Y, Xie W et al (2026) How memory management impacts LLM agents: an empirical study of experience-following behavior. In: *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, San Diego, pp 623–645. <https://doi.org/10.18653/v1/2026.acl-long.27>
114. Yao S, Zhao J, Yu D et al (2023) ReAct: synergizing reasoning and acting in language models. In: *The Eleventh International Conference on Learning Representations (ICLR 2023)*. OpenReview.net. https://openreview.net/forum?id=WE_vluYUL-X. Accessed 29 Aug 2026
115. Yu Y, Yao L, Xie Y et al (2026) Agentic memory: learning unified long-term and short-term memory management for large language model agents. In: *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, San Diego, pp 21457–21483. <https://doi.org/10.18653/v1/2026.acl-long.981>
116. Zhong Z, Wu Z, Manning CD et al (2023) MQuAKE: assessing knowledge editing in language models via multi-hop questions. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, pp 15686–15702. <https://doi.org/10.18653/v1/2023.emnlp-main.971>

Figures

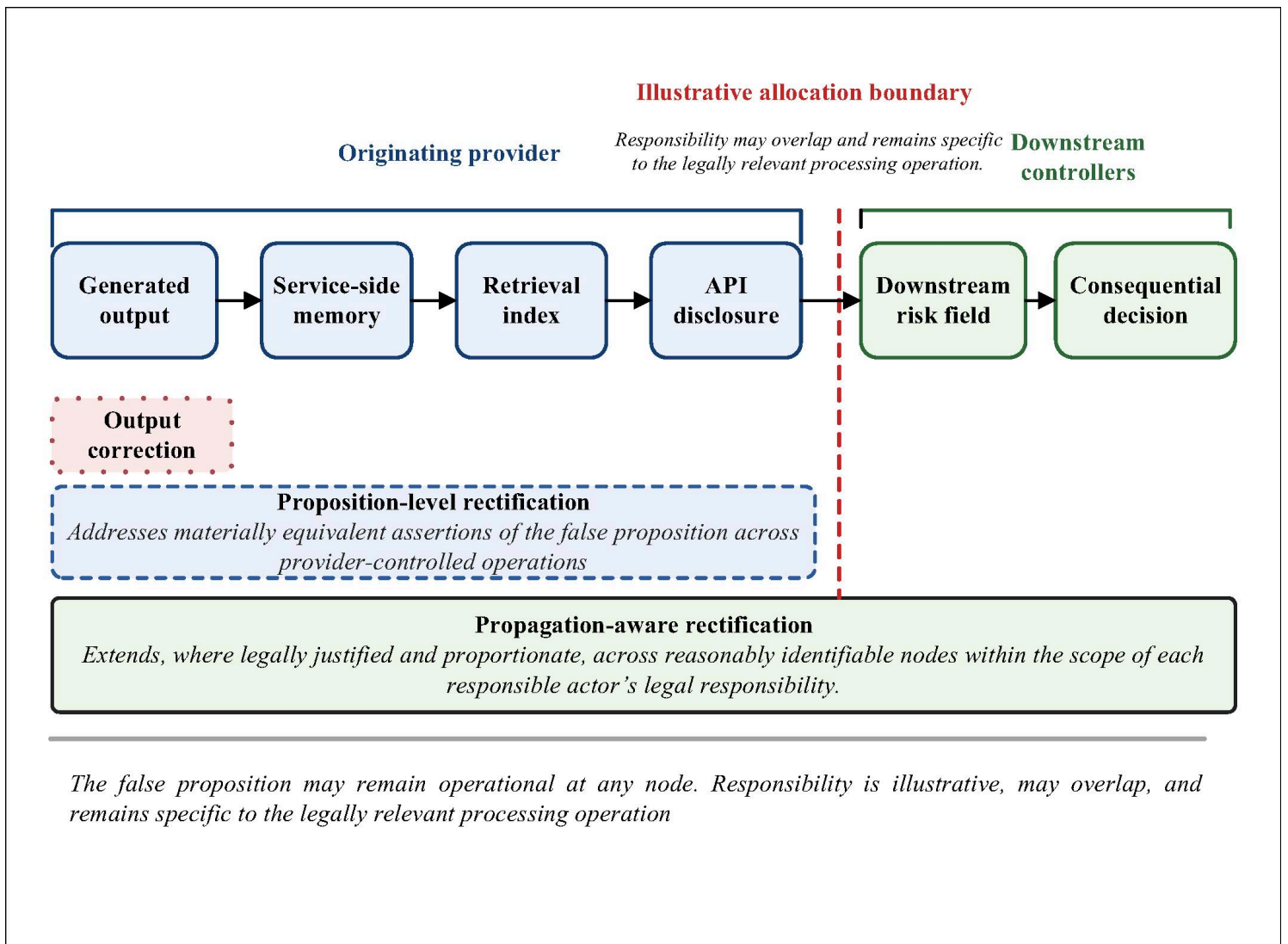


Figure 1

Propagation-aware rectification model.

Schematic representation of the movement of a false criminal allegation through six remedially relevant nodes: generated output, service-side memory, retrieval index, API disclosure, downstream risk field and consequential decision. The figure distinguishes output correction, proposition-level rectification and propagation-aware rectification. Responsibility boundaries are illustrative and remain operation-specific.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Tables.docx.docx](#)