Mitigating AI Hallucination: Balancing Automation and Human Validation

Meenakshi Achalia
Sriram Menon Koottala
Gopika K
Maria Martin
Tata IISc Medical School Foundation
(quality.manager@iiscmedicalschoolfoundation.org)
(junior.executive@iiscmedicalschoolfoundation.org)
(intern6@iiscmedicalschoolfoundation.org)
(intern7@iiscmedicalschoolfoundation.org)

ISBN: 978-1-943295-26-5

Artificial Intelligence and machine learning help researchers and professionals write and analyse efficiently, but unsupervised use can produce fabricated data, false citations, and misleading analysis. This phenomenon, termed hallucination, has been reported across academic, legal, government, and healthcare sectors. Such incidents risk financial, reputational, and ethical harm when outputs are accepted without verification. This paper presents documented cases of AI hallucination in national and international contexts and proposes a structured two level validation framework in which AI systems generate initial outputs that are subsequently reviewed by qualified human experts. Integrating this process into institutional governance and policy ensures accountability, traceability, and transparency while preserving the efficiency benefits of AI assisted work.

Keywords: Artificial Intelligence, Hallucination, Automation, Human Validation, Ethics, Governance

1. Introduction

The rapid integration of artificial intelligence into research, healthcare, finance, and governance has revolutionised efficiency but introduced new risks. Large language models and predictive algorithms are now routinely used to draft research reports, generate documentation, interpret diagnostic data, and detect fraud. These systems, however, are probabilistic text and pattern generators, not factual engines. They can produce content that sounds authoritative yet is entirely false a phenomenon known as AI hallucination.

Hallucination in AI is more than an academic curiosity; it has tangible social, financial, and legal implications. Governments have issued flawed reports, academics have submitted unverified data to official inquiries, and algorithms have made biased or incorrect decisions about citizens' welfare eligibility. In healthcare, generative and reasoning models have presented fabricated diagnostic evidence as fact.

This paper analyses six cases of AI hallucination across diverse domains government reporting, academic research, welfare automation, legal documentation, and healthcare diagnostics. Together, these examples demonstrate the urgent need for a human validated operational framework that embeds oversight, traceability, and accountability into every stage of AI assisted work.

2. Documented Incidents of AI Hallucination

Case 1: Fabricated References in a National Welfare Compliance Review (2024–2025)

In December 2024, a major consulting firm was commissioned by a national employment and welfare department to conduct an independent review of a Targeted Compliance Framework a digital system that automatically issues penalties when job seekers fail to meet government mandated mutual obligation requirements. The contract, valued at approximately $440,000, sought to examine whether the framework's design and IT logic reflected the intent of the underlying legislation and whether systemic errors or algorithmic biases existed.

When the report was first published in July 2025, it appeared to offer a comprehensive technical assessment, identifying problems such as a lack of "traceability" between the policy rules embedded in the IT system and the legislative provisions they were meant to implement. However, within weeks of release, academic reviewers discovered multiple serious errors throughout the document specifically, non-existent legal citations, fabricated court cases, invented professor names, and reference lists that included entirely fictional publications.

An independent scholar, reviewing the document's footnotes, traced the anomalies to the use of Azure Open AI gpt-4o, a large-language-model tool integrated into the consulting firm's internal workflow through the client department's cloud tenancy. The AI system had been employed to draft background text and generate reference lists that analysts were meant to verify later. However, due to deadline pressures and internal review bottlenecks, this verification stage never occurred.

When the issue became public, the firm issued a revised version of the report, adding a brief statement in its appendix acknowledging the use of "a generative artificial intelligence (AI) large language model based tool chain." It stated that corrections had been made to "a small number of references and footnotes" and asserted that these amendments did not change the report's overall findings. Nonetheless, detailed re analysis by academic observers revealed that the updated version replaced one hallucinated citation with multiple new, unverifiable ones, suggesting that the evidentiary basis of the analysis had never been properly grounded in authentic literature.

Further investigation found that one reference cited a non-existent court decision in a welfare-related judicial case, while others attributed research to professors affiliated with the University of Sydney and Lund University both institutions confirming no such papers or scholars existed. These discoveries undermined confidence in the report's methodology and raised public questions about the firm's internal quality controls.

As a consequence, the national department withheld the final instalment of the contract payment, and the consulting firm voluntarily refunded part of the project fee. Although the department maintained that the recommendations of the report remained valid, it publicly acknowledged the lapse in verification and reaffirmed its commitment to rigorous data-integrity standards for future contracted research.

This case demonstrates how AI hallucination can propagate unchecked through institutional reporting pipelines when efficiency is prioritised over validation. The use of gpt-4o as a drafting assistant without explicit traceability protocols meant that staff could not identify which sentences originated from AI, which were edited by humans, or which had been fully verified. It also illustrates how reputational harm extends beyond the consultant implicating the commissioning government body and eroding public confidence in official reports.

Operationally, this incident underscores three lessons:

- Human validation cannot be optional; every AI-assisted deliverable must undergo line-by-line fact checking.
- Traceability is essential, requiring systems that tag and archive the provenance of each paragraph and reference.
- Accountability must be contractually defined, so that the use of generative tools is disclosed at project inception, not post-publication.

Without such safeguards, even well-intentioned automation risks transforming legitimate consultancy outputs into reputational liabilities and distorting the public record.

Case 2: False Allegations in a Parliamentary Inquiry Submission (2024)

In mid 2024, a parliamentary committee examining governance in the consulting industry received a written submission from a group of academics calling for stronger accountability measures. One of the contributors, seeking to streamline the preparation process, used Google Bard, a conversational generative AI model, to create illustrative case studies and examples of alleged corporate misconduct.

The output appeared convincing. Bard produced detailed narratives of scandals, disciplinary actions, and audit failures complete with specific firm names, senior partners, and years of occurrence. These descriptions were entirely fabricated. The model "hallucinated" a set of events that had never taken place, yet they were included verbatim in the submission, along with references to non-existent court cases and fabricated internal documents.

Among the spurious examples were allegations of a so-called "wage-theft scandal" involving a large consulting firm and a banking audit said to have resulted in partner dismissals events later confirmed never to have occurred. Other claims wrongly accused another firm of auditing a major national bank during a financial planning controversy and being sued for negligence in connection with a construction company collapse. None of these assertions were grounded in fact; all were AI generated fabrications.

The submission, granted parliamentary privilege, was uploaded to the official inquiry website, where it could be quoted by the media without legal risk. Within days, affected organisations lodged formal complaints, warning that the dissemination of false material under parliamentary protection had tarnished reputations and misled policymakers.

The committee initiated a review and concluded that the use of Google Bard had indeed generated "fictional examples and inaccurate references." It issued a formal statement noting that "emerging tools within the artificial intelligence space, whilst appearing to reduce workload, may present serious risks to the integrity and accuracy of work if they are not adequately understood or applied in combination with rigorous fact checking."

The lead academic accepted sole responsibility, confirming that he had used Bard to draft sections of the submission without human cross-verification. He and the co-authors issued a written apology to the committee and to the named organisations. One university later clarified that a co-author who specialised in accounting ethics had not been given an opportunity to review the AI produced content before submission and would have objected to its inclusion.

The parliamentary inquiry treated the event as a case study in academic and procedural vulnerability. It raised questions about whether current research ethics policies, peer review norms, and institutional oversight mechanisms were equipped to manage the infiltration of unverified AI generated material into public policymaking. Legislators emphasized that academic freedom does not absolve contributors of the responsibility to ensure factual accuracy and transparency in authorship.

This case illustrates several systemic weaknesses

- Automation bias: the misplaced trust that AI outputs are accurate because they appear coherent.
- Lack of human validation: no process was in place to cross-check AI text before publication.
- Integrity dilution: parliamentary privilege inadvertently shielded falsehoods from immediate correction.

The fallout from this incident extended beyond reputational damage. It catalysed a parliamentary discussion on establishing formal AI use disclosure requirements for all future submissions and triggered academic institutions to re-evaluate internal research submission policies.

Ultimately, this case reveals how generative AI can inject misinformation directly into democratic processes, bypassing conventional editorial and peer-review filters. When outputs from systems like Google Bard are accepted without verification, the result is not merely academic error but a breakdown of procedural truth in policymaking itself.

Case 3: Algorithmic Bias and Hallucinated Decision Patterns in Public-Sector Automation (2020–2024)

Between 2020 and 2024, several government departments introduced algorithmic and artificial intelligence based tools to support decision-making in welfare, immigration, and fraud prevention systems. These tools were designed to detect anomalies, flag high risk cases, and accelerate case handling in large scale administrative operations. However, over time, evidence emerged that many of these systems exhibited patterns of bias, opacity, and hallucinated logic producing outcomes that were not fully explainable, traceable, or legally transparent.

One of the earliest examples was a visa-sorting algorithm introduced in 2020 by a central government department to help manage immigration applications. The model used a traffic light risk scoring system, where applicants were automatically classified as red, amber, or green based on perceived risk factors. Following a legal challenge by civil-society groups, it was revealed that the system had been automatically assigning red scores to applicants from certain nationalities, effectively embedding racial and geographic bias into the decision process. The department suspended the tool after acknowledging that its design might have “entrenched systemic discrimination.”

A parallel controversy developed in 2021–2024 around welfare and benefit administration. The national welfare department had adopted a machine learning fraud detection system to identify potential benefit fraud in universal credit claims. The algorithm, part of an integrated risk and intelligence service, flagged cases for further human review. However, campaigners and members of Parliament received complaints from hundreds of claimants particularly from certain Eastern European communities, whose benefits were suspended without clear explanation, leading to months of destitution and eviction proceedings.

Subsequent investigation found that these suspensions were triggered by semi-automated data analytics models, which classified cases as “high risk” based on unclear criteria. While the department insisted that nationality was not explicitly used as an input variable, it also conceded that the model was self-learning and that it was impossible to determine how it weighted different factors. In one parliamentary intervention, a legislator cited multiple cases in which claimants’ benefits were restored after lengthy appeals, with no evidence of fraud ever being established. The department’s own equality assessments, disclosed only in part, revealed that it had limited ability to test for bias due to insufficient user data.

The lack of transparency became a broader public issue. Advocacy groups such as the Public Law Project (PLP) and other legal charities filed Freedom of Information requests demanding access to model documentation, training data, and fairness analyses. In many instances, government agencies invoked national security and fraud-prevention exemptions to deny disclosure. The result was an escalating public debate about whether algorithmic systems were making unverified, self-reinforcing decisions that mimicked human reasoning without evidentiary basis a form of operational hallucination.

The controversy expanded to include the Home Office, which used a separate algorithmic model to flag potentially “sham” marriages. An internal equality assessment revealed that marriages involving citizens of Greece, Albania, Bulgaria, and Romania were being disproportionately referred for investigation. The report justified this disparity as “indirectly discriminatory but proportionate to overall objectives.” Civil rights advocates warned that such reasoning institutionalized bias by allowing models to treat correlation as causation, without verifying the validity of the risk signal.

In response to mounting criticism, the government’s Centre for Data Ethics and Innovation (CDEI) later renamed the Responsible Technology Adoption Unit published a 2020 report acknowledging that public-sector AI systems had “entrenched or amplified historic biases, or even created new forms of unfairness.” In 2021, the CDEI introduced an Algorithmic Transparency Recording Standard, requiring government departments to publish details about the AI tools they deployed, including their purpose, training data, and governance mechanisms. Despite this, by 2024 only nine AI systems were listed on the official transparency register, none belonging to the welfare or immigration departments responsible for the most contentious applications.

Further controversy arose when a fraud detection system for Universal Credit advance claims, deployed by the welfare department in 2021, was found to lack any published fairness evaluation. While the department asserted that internal reviews showed “no immediate concerns of discrimination,” it refused to release those assessments, arguing that disclosure would enable fraudsters to manipulate the model. The Public Law Project has since compiled its own independent registry of 55 government-operated algorithmic tools, many of which lack any public documentation of bias mitigation or human-oversight mechanisms.

Additional concerns extended to law enforcement agencies, which implemented live facial recognition systems to identify individuals on watch lists. A 2024 review by the National Physical Laboratory found that these systems exhibited “very low error rates” overall but, when sensitivity settings were adjusted to catch more individuals, they falsely identified five times more Black people than white people. The same bias amplification pattern was observed in predictive policing programs attempting to forecast crime hotspots and “high-harm offenders.”

Regulatory oversight bodies have taken mixed positions. The Information Commissioner’s Office (ICO) reviewed several of these tools, concluding that most were deployed “responsibly and with sufficient human intervention to avoid harm,” but also cautioned that live facial-recognition deployments remained “potentially intrusive” and required on going monitoring. Meanwhile, public-interest lawyers argue that the cumulative evidence points to a wider governance failure: AI and algorithmic systems were being introduced faster than their legal, ethical, or equality frameworks could adapt.

This case illustrates a crucial dimension of AI hallucination: not the invention of false text or data, but the emergence of spurious or biased decision patterns within machine-learning models that behave as though they are reasoning agents without true causal understanding. The operational outcomes wrongly suspended benefits, discriminatory referrals, or disproportionate surveillance represent algorithmic hallucinations embedded in bureaucratic logic.

Key Lessons from this Episode Include

- Transparency must be mandatory, not voluntary. Public institutions cannot rely on internal audits alone; all algorithmic tools must be registered, documented, and subjected to external scrutiny.
- Bias monitoring requires continual data disclosure. Departments must publish equality-impact summaries and enable independent replication of fairness tests.
- Human accountability must remain central. Automation should never suspend welfare, immigration, or civil rights without direct and reviewable human decision-making.

The broader takeaway from these developments is that the integrity of AI-driven governance depends not merely on model accuracy but on traceable human oversight. Without it, the promise of efficiency transforms into a systemic risk where algorithmic hallucinations, embedded deep within decision infrastructure, replicate and scale the very inequities that automation was meant to eliminate.

Case 4 – The Dutch Algorithmic Welfare Scandals: From SyRI to the Child-Benefits Crisis (2013–2021)

The Netherlands offers one of the most striking examples of how algorithmic decision-making can drift into large-scale discrimination and systemic injustice under the guise of efficiency. Two interconnected scandals the suspension of the SyRI (System Risk Indication) program in 2020 and the subsequent Child-Benefits Affair (Toeslagenaffaire) that peaked in 2021- demonstrate how automated systems can “hallucinate” patterns of fraud where none exist, producing real world harm to thousands of citizens.

The SyRI system was developed by the Dutch Ministry of Social Affairs and Employment as an automated risk-profiling model to detect potential welfare and tax fraud. Built over the previous decade, SyRI aggregated data from multiple government databases including employment history, debt records, education, and housing details and applied a secret risk algorithm to flag individuals or households considered “high risk.” It was deployed primarily in low-income neighbourhoods as part of a national anti fraud strategy.

By design, SyRI was opaque. Its algorithmic logic, data weighting, and error rates were not disclosed to the public, legislators, or the affected citizens. Civil rights advocates argued that the system effectively spied on the poor, generating suspicion without individual cause and disproportionately targeting vulnerable groups and migrant communities. In 2020, a coalition of privacy organisations, welfare rights groups, and the country’s largest trade union filed a legal challenge, asserting that SyRI violated the European Convention on Human Rights (ECHR) by failing to maintain a fair balance between fraud prevention and the right to privacy.

The District Court of The Hague ruled in February 2020 that SyRI was unlawful, finding that it lacked adequate transparency, safeguards, and proportionality. The judgment concluded that the system’s secrecy made it impossible for citizens to contest decisions or understand why they were targeted, effectively breaching human rights law. The court’s decision became one of the first in the world to halt a state-run AI surveillance system on human-rights grounds. The United Nations Special Rapporteur on Extreme Poverty, Philip Alston, called it “a clear victory for all those concerned about the digital welfare state,” warning that unregulated algorithmic governance risked “sleepwalking into a digital dystopia.”

While the SyRI program was formally suspended, a second, related scandal emerged within the Dutch Tax and Customs Administration. Between 2013 and 2019, the agency employed another machine-learning system to detect childcare benefit fraud. This algorithm flagged tens of thousands of families for investigation, often on the basis of dual nationality, postal code, or minor documentation discrepancies. These “risk indicators” akin to digital stereotypes functioned as algorithmic hallucinations, mistaking correlation for causation and inflating suspicion into assumed guilt.

The consequences were devastating. More than 20,000 families were wrongly accused of fraud, ordered to repay tens of thousands of euros, and cut off from financial support. Many were low income, dual national, or immigrant households. Parents lost jobs, housing, and even custody of their children. Internal audits later confirmed that nationality had been an active variable in the model’s training data, embedding direct discrimination into the algorithmic pipeline.

By the time the scandal became public in late 2019, the social and political damage was irreparable. A parliamentary investigation declared it a “systemic failure of the rule of law.” The Dutch government accepted responsibility, and in January 2021, the entire Cabinet of Prime Minister Mark Rutte resigned en masse, acknowledging institutional bias and governance collapse.

In the wake of the crisis, the Netherlands implemented sweeping reforms to rebuild public trust. It launched the Algorithm Register of the Netherlands, a public repository requiring agencies to document all algorithmic systems in use, including their purpose, data sources, and oversight mechanisms. The government also strengthened its Algorithmic Accountability Framework, mandating human rights and bias impact assessments for any AI tools that influence public services. These measures directly informed the European Union’s forthcoming AI Act, which now classifies welfare, education, and immigration algorithms as “high-risk systems” subject to strict transparency and supervision rules.

The Dutch experience encapsulates two critical dimensions of AI hallucination:

- **Data-driven hallucination :** where self-learning algorithms misperceive patterns of fraud or risk based on biased data inputs; and
- **Governance hallucination:** where institutions mistake algorithmic outputs for objective truth, abandoning critical human judgment.

Operationally, the Dutch scandals underline the necessity of traceable, explainable, and contestable AI systems. They show that algorithmic governance without proportionality, transparency, and independent validation can replicate and even amplify the injustices it was designed to prevent. The legal precedent established by the Dutch courts now serves as a global warning that efficiency cannot justify opacity, and that the digital welfare state must remain subordinate to the principles of human rights and due process.

Case 5: Legal Hallucinations in the U.S. Judicial System (2023)

In May 2023, a case before a New York federal court revealed a startling example of legal hallucination produced by generative artificial intelligence. Two attorneys representing a client in a personal injury lawsuit submitted a legal brief containing multiple citations that appeared genuine but were entirely fabricated. The lawyers had used Chatgpt, a conversational large-language model, to draft sections of the court filing and to generate supporting case law references.

At first glance, the citations were formatted correctly and attributed to real judges, courts, and legal reporters. However, upon review by the opposing counsel and the presiding judge, none of the cases could be located in any legal database. Further investigation revealed that Chatgpt had "hallucinated" the cases inventing legal precedents, court numbers, and even plausible judicial reasoning to support the argument. The lawyers later admitted that they had asked the AI whether the cases were real and that it had confidently confirmed their authenticity.

The presiding judge issued a scathing order describing the incident as "an unprecedented circumstance of fake judicial authority" and imposed financial sanctions on the attorneys and their firm. The case became a global illustration of how generative models can fabricate highly convincing but wholly fictitious factual and procedural content.

Beyond the embarrassment for those involved, the event exposed a systemic weakness in legal and administrative workflows increasingly assisted by AI drafting tools. Legal professionals under time pressure may unknowingly treat AI generated outputs as verified references. Because systems like Chatgpt are trained on probabilistic text patterns rather than structured databases, they can generate citations that appear statistically plausible but have no real world referent.

The incident sparked an industry-wide debate about whether AI generated legal materials should be treated as confidential drafts or as "expert tools" subject to the same validation standards as human paralegals. Bar associations in the United States and United Kingdom subsequently issued advisories requiring that any use of AI-generated material in court submissions be explicitly disclosed and independently verified.

This case stands as the first widely publicised example of AI hallucination entering the judicial record, illustrating how generative systems can create false legal realities when human oversight is removed. It also reinforces the need for structured traceability protocols, where AI generated text is marked, reviewed, and attributed before entering any formal or evidentiary process.

Case 6: Diagnostic Hallucinations in Clinical AI Systems (2023–2024)

In the healthcare domain, hallucination has moved beyond language models to medical reasoning systems used in clinical decision support. A prominent example emerged between 2023 and 2024, involving Google's Med PaLM and its successor Gemini, both designed to provide expert-level medical responses to patient cases and diagnostic queries.

These models were trained on a vast corpora of medical textbooks, clinical guidelines, and anonymised case records. In internal evaluations and limited pilot programs, they demonstrated strong performance on factual recall and reasoning questions often achieving scores comparable to medical residents on licensing examinations. However, subsequent testing by external researchers and medical professionals revealed a recurrent pattern of diagnostic hallucination, in which the AI confidently produced plausible but clinically incorrect reasoning chains or misinterpreted patient symptoms.

For instance, in controlled studies at academic medical centres, Med Palm occasionally fabricated laboratory findings, cited non-existent clinical trials, or drew unsafe treatment inferences that were not supported by any underlying data. In one test case involving chest pain, the model correctly identified cardiac ischemia as a possibility but then added "supporting evidence" from an invented ECG reading and a fabricated reference to a cardiology study. Similar issues arose with Gemini, where model hallucination led to internally consistent but factually incorrect medical explanations, posing potential risks if used unsupervised.

While these systems were still under evaluation and not used for direct patient care, their performance raised ethical and operational alarms in the medical community. Clinicians and regulators highlighted that AI systems trained on incomplete or uncurated datasets can project an illusion of diagnostic certainty, masking the probabilistic nature of their reasoning. The issue became particularly concerning in telehealth and automated triage systems that rely on such models to summarise patient records or generate draft clinical notes.

Following the discovery of these inaccuracies, Google and other healthcare AI developers publicly committed to maintaining human in the loop validation, ensuring that outputs are reviewed by licensed clinicians before use. Independent assessments by medical ethicists described the phenomenon as a form of "computational overconfidence" a type of algorithmic hallucination where models not only produce false data but also justify it within a coherent narrative that appears rational to human reviewers.

These diagnostic hallucinations underscore the high stakes of AI deployment in healthcare, where errors can directly translate into harm. They demonstrate that transparency, interpretability, and verification protocols are as critical in medicine as in any scientific or governmental domain. The incident reinforced the consensus that AI should assist, not replace, clinical judgment and that all AI generated medical reasoning must remain traceable, contestable, and explicitly labelled as machine-derived.

3. Analysis of Underlying Causes

Across these cases, several consistent drivers of hallucination emerge:

- **Probabilistic Reasoning without Fact Anchors:** Generative models predict plausible sequences of words or classifications but lack internal mechanisms for truth validation.
- **Opaque Training Data**: Many systems are trained on unverified or biased data sets, embedding errors and prejudices.
- **Absence of Traceability:** Without clear demarcation of AI generated content, human reviewers cannot reliably audit information chains.
- **Automation Bias**: Users tend to trust outputs that appear fluent or statistically confident, even in the absence of corroborating evidence.
- **Institutional Over-Reliance on Efficiency:** Pressure for speed and cost reduction disincentivises deep verification, allowing hallucinated material to pass through unchecked.

4. The Risk Context in Healthcare and Research

Healthcare and scientific research are uniquely vulnerable to AI hallucination.

- In research, hallucinated citations and data can propagate misinformation into peer-reviewed literature. In clinical environments, diagnostic hallucination can lead to mistreatment and patient harm. Hospitals using AI for finance, reporting, or electronic records risk reputational damage and regulatory penalties if false or unverifiable information enters official systems.

Accreditation frameworks such as NABH, JCI, and ISO 13485 already emphasise traceability and data integrity. Their principles can be adapted to govern AI use requiring version control, author identification, and validation documentation for all AI assisted outputs.

5. Proposed Two Level Validation Framework

Level 1: AI Draft Generation

AI systems may generate structured drafts or analytical summaries to accelerate documentation, reporting, or diagnostics. All outputs must be logged with metadata identifying model version, prompt origin, and generation timestamp.

Level 2: Human Expert Verification

Human reviewers subject-matter experts or accredited auditors must verify:

- The factual accuracy of every quantitative statement,
- The authenticity of every citation and dataset, and
- The compliance of content with ethical and regulatory standards.

Each final report should maintain a trace log showing which sections were AI generated, human edited, and expert verified. This transparency protects both institutional integrity and public accountability.

6. Operational Integration and Oversight

To institutionalise this framework:

- Establish an AI Oversight Committee within every hospital or research body, responsible for approving AI-assisted documentation workflows.
- Mandate AI disclosure statements in all research papers and policy reports.
- Incorporate AI validation training into continuing professional education for researchers and clinicians.
- Ensure AI systems comply with existing standards for data security, ethics, and auditability under NABH, JCI, and ISO.

7. Discussion: Balancing Automation and Human Trust

Automation offers unparalleled productivity but must coexist with human judgment. AI hallucination represents a failure not only of algorithms but of governance. Institutions must recognise that efficiency cannot substitute for accountability. The objective is not to slow down automation but to make it reliable, explainable, and ethically defensible.

8. Conclusion

AI hallucination is an emerging cross sectoral risk that undermines the credibility of automation. The six cases documented here reveal that without structured human validation, generative and predictive systems can fabricate evidence, distort decisions, and erode institutional trust. A two level framework AI generation followed by mandatory human verification offers a practical model for mitigating these risks.

When implemented through policy, accreditation, and education, this approach ensures that AI remains a tool of augmentation, not deception, preserving scientific integrity and public confidence in the era of intelligent automation.

9. Acknowledgements

The authors thank the Tata IISc Medical School Foundation and its Clinical Administration and Quality teams for their continued support in developing ethical frameworks for responsible AI in healthcare.

10. Author Contributions

Mrs. Meenakshi Achalia conceptualised the paper and oversaw editorial integration. Dr. Sriram Menon Koottala contributed to case compilation and framework development. Ms. Gopika K and Ms. Maria Martin performed literature verification and structural editing**.**

11. Conflict of Interest

The authors declare no conflicts of interest and no financial relationships that could have influenced the content of this paper.

12. References

1. Deloitte welfare framework report (GPT-4o) – Guardian, Oct 2025.
2. KPMG academic submission & Bard hallucination – Guardian, Nov 2023.
3. UK public-sector algorithmic bias & DWP fraud tools – Guardian, Oct 2023.
4. Dutch SyRI and child-benefits scandals – Guardian, Feb 2020 + Politico recap.
5. New York ChatGPT legal-brief incident – Guardian, May 2023.
6. Clinical hallucinations (Med-PaLM / Gemini) – Guardian Science coverage, 2023–24.