

WHEN AI LIES: WHO'S LIABLE?

Adv. Vyomesh Uday Warunjikar

Abstract:

Artificial Intelligence (AI) has seamlessly integrated into daily life, particularly through Large Language Models (LLMs) like ChatGPT, Gemini and others, which now influence legal research, writing and discourse. However, these models can produce misleading or fabricated content, a phenomenon called as “Legal hallucination”, posing serious challenges in high-stakes fields like law. This paper examines the legal implications of such hallucinations in the Indian context, questioning developer liability under existing regimes such as the Consumer Protection Act and Information Technology Act. It contrasts India's legal vacuum with the European Union's AI Act, which offers a more holistic approach. The paper advocates for regulatory reform, integration with verified legal databases, and the development of hallucination detection tools to ensure accountability and mitigate risk.

Keywords: Legal hallucination; Artificial Intelligence; Developer Liability; AI Accountability; AI Regulation

Copyright © 2025 The Author(s): This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY-NC 4.0) which permits unrestricted use, distribution, and reproduction in any medium for non-commercial Use Provided the Original Author and Source Are Credited.

Introduction:

We wake to the soft hum of our AI-powered alarm, smoothly easing us into the day. Before our feet even touch the floor, algorithms have already contemplated our morning. They've filtered through the news to serve us the relevant ones as per our preference, identified potential scheduling clashes on our calendar and even lined up a podcast we didn't know we would enjoy, perhaps it knows our habits and preferences better than our closest friends. Artificial Intelligence has woven itself seamlessly into the fabric of daily life.

AI doesn't just assist our day, it practically is our day, enhancing comfort, streamlining tasks and design our life with surgical precision. Unfortunately, we have grown too comfortable with this overachiever of the modern world, however, we also have comfortably grown to idea of AI being smart, flawless and unerring. But smart doesn't mean wise, even the most sophisticated systems can misfire and stumble, thereby misinterpreting the content, bringing in different kinds of biases or take decisions which we, the makers of AI don't understand. The mistakes of AI have far-reaching consequences, deeply affecting human lives, rendering it extremely significant to understand the altruistic meaning of AI.

To truly understand Artificial Intelligence, it's imperative to define intelligence. Cambridge Learner's Dictionary defines intelligence as “the ability to learn and understand things quickly and easily”(INTELLIGENCE | English Meaning - Cambridge Dictionary, n.d.), whereas, Oxford's Learner Dictionary defines it as “the ability to learn, understand and think in a logical way about things”(Intelligence Noun - Definition, Pictures, Pronunciation and

Usage Notes | Oxford Advanced Learner's Dictionary at OxfordLearnersDictionaries.Com, n.d.). Intelligence is often recognised as the ability to learn from experience, adapt to new situations, think critically and solve problems across varied parameters. For the longest time, these abilities and skills were considered distinctly human. However, in contemporary times, machines are learning to mirror these abilities and skills. This upgrading capability of machines is the crucial and core component of Artificial Intelligence, a field focused on creating systems that can mimic tasks that require human intelligence.

AI & Its Definitions:

Artificial Intelligence (AI) is not a recent phenomenon, it has evolved over the past 70 years, guided by the work of visionary thinkers and computer scientists like Alan Turing, John McCarthy, Stuart Russell and Peter Norvig. These pioneers laid their critical intellectual foundations for what we now call AI, and their definitions of the concept reflect the era's evolving technological and philosophical perspectives.

The earliest milestone came in 1950, when British mathematician Alan Turing published his influential seminal paper, introducing what is now famously known as the Turing Test, "*Any computer that passes the Turing Test*" (Alan Turing, 1950). The test ascertains a machine's ability to imitate human-like communication convincingly enough that a human interviewer cannot distinguish it from another human. This marked a landmark in AI, introducing the concept of measuring intellect through behaviour and interaction.

Fast forward to 2007, John McCarthy, who originally coined the term "*Artificial Intelligence*" (John McCarthy et al., 1955) in 1955, gave a more structured definition that highlighted AI's dual nature as scientific and engineering, "*The science and engineering of making intelligent machines*" (John McCarthy, 2007, p. 2). His definition portrays AI not just as a theoretical concept but also as a practical and technical discipline situating AI strongly within the broader ambit of computer science and applied engineering.

By 2020, the concept of AI had bloomed and matured further, with the agent-based model gaining global acceptance in both academia and industry. This perspective is best proposed in *Artificial Intelligence: A Modern Approach* (Stuart Russell & Peter Norvig, 2021) by Stuart Russell and Peter Norvig. It defines AI as, "*A software system which perceives its environment through sensors and acts upon that environment through actuators*" (Stuart Russell & Peter Norvig, 2021, p. 34). This definition conceptualises AI as an autonomous agent that constantly interacts with its environment, making decisions and taking actions relying on sensory input. It reflects the increasingly vital and real world orientation of modern intelligent systems.

Finally, the most comprehensive and governance oriented definition is provided by the European Union's AI Act, which aims to approach both, the technological capabilities of AI and its regulatory concerns. Article 3.1 defines AI system as, "*A machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments*" (Artificial Intelligence Act, 2024). Unlike previous definitions, this definition highlights the change in trend from innovation-driven focus to risk management and accountability.

Large Language Model:

Having skimmed through the wide concept of AI, we now delve into a remarkable subset of AI that have transformed how machines process and generate human language, Large Language Models or LLMs. At their essence, LLMs are advanced algorithms trained to predict the next word in a sequence of text, much like how we might finish someone's sentence based on context (Nicholas Carlini et al., 2021). This deceptively simple task of next word prediction, results in models to learn complex patterns in language, from grammar and tone to meaning and distinction. Over time, rigorous training on vast collections of datasets, ranging from books and articles to websites, LLMs gradually develop an impressive statistical understanding of how language works and enabling them to execute wide range of tasks (NITI Aayog, 2018).

The way these models process and generate text is fundamentally iterative. LLM's like Open AI's 'Chat Generative Pre-trained Transformer' (Chat GPT), Google's 'Gemini', or Bidirectional Encoder Representations from Transformer (BERT) construct responses repetitively word by word, selecting each based on the probability it fits with the preceding words. It allows them to create responses that feel natural and conversational, resembling human dialogue. Under the hood, most LLMs rely on a transformer architecture, a breakthrough that allows them to handle data effectively. During training, they consume huge volumes of text in a self-supervised fashion, thereby learning patterns without needing explicitly compartmentalised data. Once trained, these models can be guided with prompts or questions to handle specific tasks (NITI Aayog, 2018).

Hallucinations:

Chess has long been considered as a benchmark for human intelligence. Its complexity demands careful logical reasoning, strategic planning and deep focus, qualities often representing superior intellectual capability. Although, a popular YouTube video reveals an unexpected twist, wherein ChatGPT, when asked to play a simple game of chess, performs remarkably poorly by playing illegal moves. Despite being given clear instructions and access to the rules, it makes several gross errors, like moving pieces that had already been captured or teleporting a bishop across the board in a single move.

ChatGPT's well known chess blunder, offers more than just a funny moment, it exposes a core limitation of LLMs. Unlike dedicated chess engines like Stockfish, which calculate legal moves using rigid, rule based logic, LLMs like ChatGPT generate responses by predicting the most likely next word or sentence based on patterns in text. They excel at sounding fluent and confident, but that doesn't mean they are always right. In fact, when precision and logical consistency matter, as in chess, their lack of real comprehension becomes obvious.

Furthermore, they can also be biased and produce inaccurate or harmful outputs depending on the size and quality of the training dataset. As a result, they can sometimes generate responses that are seem accurate but are factually incorrect, an issue known as "*AI hallucination*". It is the model's attempt to fill in the vacuum in knowledge or data, with a speculation based on the patterns it has learned during training (Muhammad Usman Hadi et al., 2024). It raises concerns in high-stakes fields like law, medicine, or governance. The model from its huge training dataset cannot tell a difference a Supreme Court judge's opinion and a first year undergraduate LL.B student's opinion, thereby creating "*Legal Hallucination*".

AI Hallucination is a very broad concept, not everything can be addressed in a research paper this length, therefore boundaries ought to have been created. Unlike general errors, legal hallucinations can mislead legal arguments, cause ethical violations, or result in wrong and inaccurate decisions. Often undetectable to non-experts, they risk eroding public trust in justice systems and pose serious professional and societal consequences when relied upon blindly. Unfortunately, very few countries have been able to tame this digital wildfire. This paper will be confined to legal hallucinations and its regulatory concern along with its liability and accountability in the Indian scenario.

In courtrooms and legislative halls, accuracy is not an option, it's a fundamental duty. For lawyers, precision is a matter of professional ethics. For judges, it's imperative to deliver justice. And now, as AI becomes a pertinent player in legal research, journalism, content creation and public debate, accuracy is evolving into a legal and moral imperative for developers as well. As AI has started usurping the roles once reserved for human experts, a crucial question emerges as to who is accountable when these systems get it wrong? When AI fabricates citations, misinterprets law, or presents fictional claims with evident authority, who bears the blame?

Accountability of Legal Hallucinations in Indian Context:

Traditionally, the burden of authentication has fallen on the user. Lawyers, for example, are ethically bound to crosscheck all facts and legal references prior to submitting them in court. Judges, thereafter, are expected to examine the evidence and arguments with an independent judicious mind. But AI appearing to be dominant and authoritative complicates this dynamic. Its responses sound persuasive even when they're inaccurate, making it harder to tell truth from the fallacy. This blurring of lines increases the risk of accidental misinformation.

However, it's neither fair nor pragmatic to place all the responsibility on the end-user. Developers of general-purpose AI, especially those offering tools for use in legal research, policy making, arguments and documents drafting, judgement writing hold a significant share of the responsibility. As providers of services (digital), they may fall within the ambit of product and service liability regimes. Legal concepts like negligence, duty of care and deficiency in services may be applicable here, which can be invoked by filing a suit for damages in torts.

The use of AI in the legal field heightens the stake, leading to invalid filings, wrong citations, non-existent judgements, or misworded statutes. If an AI platform is used as a professional grade legal assistant, developers can't simply wash their hands of what the system produces.

The Indian legal definitions of liability are anchored in the idea of a "person", either natural or juristic. An AI system, which lacks consciousness or *mens rea* cannot be held liable, however, its developers can be brought under the ambit of definitions across varied statutes. Firstly, section 2(7) of the Consumer Protection Act, 2019, defines a consumer as,

"hires or avails of any service for a consideration which has been paid or promised or partly paid and partly promised, or under any system of deferred payment and includes any beneficiary of such service other than the person who hires or avails of the services for consideration paid or promised, or partly paid and partly promised, or under any system of deferred payment, when such services are availed of with the approval of the first

mentioned person, but does not include a person who avails of such service for any commercial purpose”(Consumer Protection Act, 2019, n.d.)

LLM’s like Chat GPT use a ‘Freemium’ model, wherein the initially for certain time or certain prompts the usage of the services is free and they won’t charge until the expiry of the same. Applications like Spotify, Gmail, or Youtube also follow the similar freemium model, highlighting its use worldwide. Therefore, it can fall under the ambit of Consumer Protection Act, 2019, thereby allowing users to file a consumer complaint in case of deficiency of services. However, it applies only once you upgrade yourself to the premium version, and not the initial free version, thereby creating a partial legal vacuum. The concept of a “complaint” or a “consumer” may not encompass users of free or publicly accessible AI systems. Furthermore, if existing law specific research platforms like SCC OnLine, Manupatra or LexisNexis equip themselves with a much advance LLM then they can be brought under the ambit of consumer commissions.

Secondly, beyond civil liability, AI developers may also find themselves navigating digital content laws, specifically under the Information Technology Act, 2000. Section 66A of the enactment lays down the punishment for transmission of false news or misleading information online causing harm. It is read as follows: *“Any person who sends, by means of a computer resource or a communication device,— (b) any information which he knows to be false, but for the purpose of causing annoyance, inconvenience, danger, obstruction, insult, injury, criminal intimidation, enmity, hatred or ill will, persistently by making use of such computer resource or a communication device”*(Information Technology Act, 2000, n.d.)

However, the provision requires *mens rea*, i.e., motive or intention, absolving the developers of any criminal liability. The hallucinations are not a result of any particular agenda or programming of the developer, it is a drawback of the technology, thereby virtually creating a legal vacuum.

In comparison, the European Union’s AI Act (2024) stands out as a robust and progressive legislative model that not only addresses but also resolved many of the regulatory gaps. Rather than focusing solely on commercial usage, the Act introduces a function-based liability framework thereby including free services as well. This means accountability is tied to the role of individuals or organisations involved in developing or deploying AI systems. Article 3(2) of the EU AI Act defines a “*provider*” as:

“a natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge.”(Artificial Intelligence Act, 2024)

This inclusive and broad definition brings even non-commercial providers within regulatory scope. Furthermore, the Act defines a “*general-purpose AI model*” as:

“an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of

downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market.”(Artificial Intelligence Act, 2024)

These definitions creates a regulatory system which covers systems like Chat GPT, Gemini, and other similar models. In addition to this, Article 15 of the EU AI Act introduces obligations and mandates for the AI developers to follow and complete the threshold of transparency, technical documentation, and mitigation of risks. This overcomes the issue of proving *mens rea* or intent, making the developer liable of its harmful shortcomings. The European Union Artificial Intelligence Act, 2024, directly addresses the accountability gaps prevailing in Indian Statute, thereby offering a compelling regulatory statute. India can definitely benefit from adopting a similar framework which ensures ethical and equitable AI governance.

Additionally, the issue is not only legal but also technical. Without standard benchmarks, neither judiciary nor lawyers have a clear measure to determine the AI systems smooth functioning. There can be benchmark much like the food industry’s ‘FSSAI’ which can define metrics for accuracy and reliability. These standards must be tailored as per the context, strict for high stakes fields like healthcare and legal ecosystems and lenient for creative writing wherein the risk is much lower.

Another approach and perhaps a second layer solution is to integrate Retrieval-Augmented Generation (RAG) frameworks. By connecting AI models with external, authentic data sources providing legal database such as Supreme Court Reporters, SCC OnLine, All India Reporters, LexisNexis, or Manupatra. Legal AI systems could be trained based on the authentic training data from the legal databases. This will greatly reduce the risk of legal hallucinations and helps anchor AI outputs in verifiable data.

Thirdly, developing tools to detect and rectify the fallacies when AI starts to hallucinate is crucial. These detectors could use logical consistency checks, statistical analyses, and verification against authoritative sources to flag inaccurate responses. There exist certain software such as Pythia, Galileo, or Cleanlab which certify and validate the data produced by AI. However, they aren’t remotely related to legal field, although, that day is not far enough when such a tool will be created for the purpose of authentication of legal data.

Conclusion:

As artificial intelligence becomes more deeply woven into the fabric of the legal profession, the discourse is shifting from convenience to responsibility. What were once seen as harmless malfunctions, like legal hallucinations, now turn into a critical issue, such as the growing tension between technological innovation and its correlative necessity for accountability. India’s legal system, while strong in its traditional approach to liability, has yet to fully adapt to the distinct challenges posed by these AI models. These systems not only assist with legal tasks but also shape how legal information is accessed, interpreted and even applied. Without the appropriate regulations, there’s an altruistic risk of eroding trust in legal processes.

To address this, India must move toward a more proactive and progressive regulatory framework. This implies designing rules that are sensitive to context, setting clear benchmarks for reliability and ensuring that AI systems are trained on verified, authoritative legal dataset. Striking this balance between stimulating technological innovation and upholding the integrity of justice is both urgent and necessary.

As AI starts to influence not only how law is practiced but how it is perceived and understood, constructing a legally sound and ethically grounded approach to AI regulation is no longer a matter of choice. It is an essential step forward. To tame the red bull of AI, we need more than reins, we need rails, routes and rules.

Cite This Article:

Adv. Warunjikar V.U. (2025). *When AI lies: Who's liable?* In **Educreator Research Journal: Vol. XII (Issue IV)**, pp. 43–49.