engineering proceedings

MDPI

Proceeding Paper

AI-Generated Inference as Algorithmic Evidence: A Systems Framework for Self-Incrimination and Constitutional Protection †

Sunkireddy Surya Prakash Reddy and Mandeep Kaur *

School of Law, Mahindra University, Hyderabad 500043, Telangana, India; sl25plaw002@mahindrauniversity.edu.in

* Correspondence: mandeep.kaur@mahindrauniversity.edu.in

† Presented at the 8th International Global Conference Series on ICT Integration in Technical Education & Smart Society, Aizuwakamatsu City, Japan, 27–29 July 2026.

Abstract

Artificial intelligence (AI) systems are increasingly transforming criminal investigation by moving beyond the retrieval or matching of existing evidence toward the generation of inferences about identity, conduct, intention, knowledge, and other aspects of human behaviour. Facial-recognition systems, predictive models, machine-learning classifiers, and multimodal foundation models can process heterogeneous digital traces and generate conclusions that may subsequently be presented as evidence in criminal proceedings. This development creates a distinctive constitutional problem because an AI-generated inference may reconstruct aspects of an accused person's mental state without the accused having voluntarily communicated that information. This paper examines this problem through the concept of AI-generated algorithmic evidence and develops a systems-oriented constitutional framework for determining when such inference should be treated as testimonial evidence under Article 20(3) of the Constitution of India. The paper distinguishes between the computational inputs used by an AI system and the inferential output generated from those inputs. It argues that the constitutional character of evidence should not be determined solely by the nature or lawfulness of the underlying data. Instead, attention must also be directed to the type of claim generated by the computational system. An inference concerning the accused's intentions, beliefs, knowledge, motives, or mental state presents a different constitutional problem from an inference concerning an external physical characteristic. Building on *Kathi Kalu Oghad* and *Selvi*, the paper proposes an inference-sensitive two-stage framework. The first stage determines whether an AI system has generated an assertion about the mental content of the accused without voluntary testimonial participation; where it has, Article 20(3) operates as a categorical constitutional restriction. The second stage applies Article 21 proportionality and a requirement of meaningful challengeability to AI-generated inferences concerning external facts. The framework further incorporates principles of explainability, algorithmic accountability, auditability, and challengeability to address the evidentiary consequences of probabilistic and potentially opaque AI systems. By treating the algorithmic system as an inferential producer rather than merely a neutral evidentiary instrument, the paper identifies a constitutional category that existing testimonial–physical distinctions do not adequately capture. The proposed approach provides courts with a structured method for evaluating AI-generated evidence while preserving the constitutional protection of mental autonomy and preventing computational authority from transforming unverified algorithmic inference into unquestionable fact.

Check for updates

Academic Editors: Debopriyo Roy, George Fragulis and Peter Ilic

Published: 4 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: algorithmic evidence; artificial intelligence; AI-generated inference; self-incrimination; explainable AI; Article 20(3); constitutional law; algorithmic accountability

1. Introduction

Criminal investigation is increasingly becoming a computational process in which artificial intelligence systems do more than retrieve, match, or organize existing evidence. Facial-recognition systems can identify persons from crowd footage [1], predictive models can rank individuals according to estimated risk [2], and computational inference systems can combine movement, association, communication, and other digital traces to generate conclusions about conduct and intent [3]. These outputs may subsequently enter judicial proceedings as a facial match, probability score, behavioural assessment, reconstructed timeline, or other computational finding presented as evidence concerning the accused. The emergence of such systems creates a constitutional question that existing evidentiary categories were not designed to address: when an AI system generates a conclusion about a person's identity, conduct, intention, or mental state that the person never communicated, what is the constitutional character of that inference?

This paper conceptualizes this problem through the notion of AI-generated algorithmic evidence. An algorithmic system can occupy a function analogous to that of an evidentiary witness when it generates an assertion about the accused rather than merely retrieving or measuring an already existing characteristic. Unlike a conventional witness, however, the computational system cannot take an oath, cannot meaningfully be cross-examined, and may generate conclusions through probabilistic relationships that are difficult even for system developers to reconstruct. The distinction is particularly significant where the system reconstructs an accused person's mental state from distributed digital traces. Existing evidence doctrine was developed around relatively identifiable sources of evidence—human testimony and physical evidence—and therefore does not map easily onto an inferential system that transforms data into a new assertion about a person's mind.

Article 20(3) of the Constitution of India provides that no accused person shall be compelled to be a witness against himself [4]. In *State of Bombay vs. Kathi Kalu Oghad*, the Supreme Court established the principal distinction governing this protection: compelling a person to communicate the contents of the mind is prohibited, whereas requiring physical evidence that exists independently of the person's will is not [5]. *Selvi vs. State of Karnataka* subsequently reinforced the protection against techniques that obtain involuntary access to the conscious contents of the mind [6]. This constitutional distinction remains fundamental, but AI-generated inference exposes a category that does not fit comfortably within either side of the traditional divide. An AI system may receive data that was not voluntarily provided as testimony and then transform those data into a computational assertion about the accused's intention, knowledge, belief, or motive.

Consider an AI system that combines location information, message metadata, transaction records, communication patterns, and other digital traces and concludes that an individual intended to commit a particular offence. The underlying information may have been automatically collected or lawfully obtained without the individual ever articulating the alleged intention. The resulting output is therefore neither conventional testimonial communication nor physical evidence in the traditional sense. It is a generated inference about mental content. The constitutional difficulty lies precisely in this transformation: the system does not obtain a statement from the accused, but computationally reconstructs a statement about what the accused allegedly thought or intended.

https://doi.org/10.3390/engproc2026143090

This distinction becomes increasingly important as AI systems evolve from conventional pattern-matching tools toward machine-learning and foundation-model architectures. Earlier forensic technologies primarily performed retrieval, comparison, and measurement tasks that could conceptually be reproduced by human analysts, although often more slowly. Machine-learning systems can instead estimate probabilities for previously unseen instances and generate inferences from statistical relationships within data [7]. Foundation models extend this capability by processing multiple modalities and supporting a broad range of downstream tasks [8]. Large language and multimodal models can produce coherent conclusions whose internal reasoning may not be fully reconstructable from the model's parameters or output [9]. Consequently, the evidentiary question is no longer limited to whether a computer has accurately retrieved information; it also concerns how a computational system transforms heterogeneous data into a new assertion about an individual.

The problem is intensified by algorithmic opacity and automation bias. Algorithmic accountability scholarship has demonstrated that complex computational systems may remain difficult to understand even where source code is available, while the proprietary nature of training data and model development can further restrict independent examination [10–12]. At the same time, computational outputs can acquire an appearance of precision and objectivity that encourages users to place excessive confidence in automated recommendations [13,14]. In a criminal proceeding, this combination of inferential complexity and perceived authority creates a significant evidentiary concern: an algorithmic conclusion about a person's mental state may be treated as fact even though neither the accused nor an independently examinable witness can explain or challenge the basis from which the conclusion was generated.

The constitutional problem therefore requires a distinction between the input layer of an AI system and its inferential output layer. The fact that the input consists of physical, transactional, or digital evidence does not necessarily determine the constitutional character of the output. A system may begin with lawfully obtained external data but generate a conclusion concerning an individual's internal mental state. The paper argues that this transformation must be examined independently because otherwise the lawful acquisition of underlying data could effectively authorize computational reconstruction of the accused's mind.

The paper accordingly develops an inference-sensitive constitutional framework for AI-generated evidence. The proposed approach distinguishes between AI-generated assertions concerning mental content and those concerning external facts. Where an AI system generates an inference about the accused's intentions, beliefs, knowledge, motives, or other mental states without voluntary testimonial participation, the paper argues that Article 20(3) should apply. Where the system instead generates an inference concerning an external fact, the constitutional analysis shifts to Article 21, including proportionality and meaningful challengeability.

The framework also recognizes that challengeability is a necessary component of computational evidence governance. Explainability mechanisms may help identify relationships between inputs and outputs, but a post hoc explanation does not necessarily establish that the underlying inference is correct [15,16]. Therefore, transparency alone cannot resolve the constitutional problem. A judicially usable AI-evidence framework must consider whether the inference can actually be examined, contested, and meaningfully challenged by the defence.

The paper thus positions AI-generated inference as a distinct algorithmic evidentiary process rather than treating AI output as automatically equivalent to either documentary evidence or conventional expert testimony. The central contribution is an inference-sensitive reading of Article 20(3), supported by a two-stage framework that separates the protection

of mental content from the proportionality and challengeability requirements applicable to external factual inference. This approach seeks to preserve the constitutional purpose of the privilege while adapting its application to computational systems capable of generating claims about the human mind without direct human testimony.

2. How Artificial Intelligence Generates Evidence

2.1. From Forensic Tool to Inferential Author

Computational investigation is nothing new, but the kind of work the computation has done in forensic analysis has changed in kind. Earlier forensic tools did work that human analysts could have done themselves, only slower: indexing files, matching fingerprints, comparing records with a database. This is all retrieval. Machine learning models do not stop there. They infer. Training a machine learning model on the existing data produces a probability estimation of the new instance. This probability estimation is a generated inference, not the discovery of the fact [7]. An identity match, a risk score, and a reconstruction of the chain of events are, thus, authored claims about the world, not observations of it.

Foundation models created an even bigger gap between inference and discovery. As the Stanford University assessment of their capabilities implies, these multimodal systems trained on broad data and used for multiple downstream tasks mark the transition from one type of AI to another [8]. These multimodal models combine text, images, geographical and behavioral information in one output, while the inference behind the output rests on a huge number of parameters whose interrelations are impossible even for the developers of the system to understand. Literature on large language models shows that the reasoning behind their conclusions is not available, while the output produced by such models is convincing and coherent [9]. An inference that reaches this far and cannot be reconstructed is exactly the kind the court needs to be suspicious of.

The law of evidence has always been concerned with the source of the assertion. The sources that can be checked by the law are the witness and the object of evidence. The witness is checked through cross-examination, and the object is checked through inspection. Inference through algorithmic means does not answer to either. Inside the model, the parameter is a correlation, not a reason. The training data are usually proprietary, and the conclusion of the inference is the probability association, not the discovery of the fact [10]. Literature on algorithmic accountability defines these models as the black boxes and shows that even publication of the source code does not help with the problem of the model's incomprehensibility for its creators [11,12].

In matters like credit scoring or benefits administration, this issue may lead to questions of fairness and due process [17]. In the matter of criminal trial, it leads to an even greater problem. In front of the court there appears an accuser who cannot be confronted.

2.2. Inference About the Mind, and the Limits of Explanation

The nature of the inferences made is equally important as their method. Tools that are used in the administration of justice do not just determine what someone has done; they make inferences about what that person was thinking, what he or she knew, or the disposition of the person to commit a crime; they assign a propensity, read intent from movements and contacts, and attribute mental states through text and metadata [3]. Such inferences relate to what is going on in the mind of a person, which is the very domain that the privilege against self-incrimination seeks to shield.

The field of Explainable AI has developed precisely because such inferences are hard to explain. In studies of Explainable AI, two types of models can be distinguished: those designed with interpretability built in and those in which interpretability needs to be engineered later on by either simplifying the complicated model or attributing weight to

some inputs [15]. The difference is not merely theoretical in view of the fact that the output is a description of what goes on inside a person's head. Post hoc explanation reveals, at best, the connection between certain inputs and the outcome of the process of inference but does not demonstrate that the inference was correct [16]; it does not provide the opportunity to test it the way it would be possible to test reasoning of the witness.

The appearance of authority is what the privilege should be especially wary of. Human opinion on intent is clearly labeled as such and comes with all the hallmarks of its subjectivity; a numeric output wears the cloak of a precise measurement and is prone to the documented tendency of automation bias, which makes people tend to rely too much on the result of computation and undervalue the conflicting input of humans [13,14]. Instead of eliciting mistrust in the face of opacity, it induces deference. An inference about a person's mental state that would be questioned if made by a human would be taken as factual evidence if issued by a machine. Rather than undermining the rationale behind the privilege, this makes it imperative: the more untestable an inference about a person's mind, the greater its importance.

3. Article 20(3) and the Limits of the Testimonial-Physical Divide

The privilege protects the mind, not the body. It is the distinction that the eleven-judge bench set out in *Kathi Kalu Oghad*, distinguishing the prohibition under Article 20(3) against forcing the accused to reveal what is in his mind from allowing him to provide physical evidence such as fingerprints or handwriting [4,5]; and *M.P. Sharma vs. Satish Chandra* had, prior to that decision, given the provision an expansive construction which *Kathi Kalu Oghad* has confined since [18]. This is a sensible distinction, and one that is robust enough to deal with the cases for which it was devised. A sample exists regardless of the accused's cooperation; a confession does not.

What *Selvi vs. State of Karnataka* adds is the reason for drawing the line where the line is drawn. Narco-analysis, polygraphy, and brain-mapping violate Article 20(3) not because they compel the accused to speak, but because they obtain involuntary access to the conscious content of his mind [6]. The injury, in *Selvi*'s understanding, is involuntary access to the mind's interior, and the means of achieving that access is irrelevant. That is why the issue can be addressed by a different lens, because the inquiry will no longer be focused on the manner of compulsion and on what it compels, but on what it accesses without it.

An objection must be met before going further. Article 20(3) speaks of compulsion, and machine inference compels nothing from the accused; on a narrow reading, the provision would therefore never be engaged. The reading defended here is *Selvi*'s own. Once the Court located the injury in involuntary access to the conscious content of the mind, compulsion ceased to be the essence of the guarantee and became one way, the classic way, of committing the essential wrong: making a person the source of evidence about his own mind against his will. The element of involuntariness does not disappear from this account; it moves. It is supplied not by force applied to the accused but by the absence of any voluntary act of his directed at furnishing the mental content the system reconstructs. And if that reading is rejected, the argument relocates rather than fails, since an untestable reconstruction of the mind would still fall to be excluded under Article 21 through the second stage of the test proposed below. The better view, however, is that *Selvi* has already carried Article 20(3) past the interrogation room, and this paper takes the Court at its word [6].

Machine inference is difficult to classify within the two-category scheme, because it does not compel the accused to speak, and the data from which it draws conclusions may be physical evidence, transactional evidence, or ambient metadata. Yet, its outcome is not an attribute of the body but an inferred mental state, which does not amount to physical evidence in the sense in which *Kathi Kalu Oghad* used that term. Machine inference reaches

the content of the accused's mind while compelling nothing; and the two-category scheme cannot accommodate it.

A way out might appear obvious but must be rejected. The argument runs that if the underlying data are physical or transactional traces, lawfully obtained by the State, then the inference based on it is simply a further analysis of physical evidence, and the privilege of Article 20(3) is not triggered. This confuses the constitutional status of the input with that of the output. Physical evidence lawfully compelled according to *Kathi Kalu Oghad* will remain physical evidence however much it is analyzed subsequently; but an inference from that evidence, and other data collected, to the accused's intent to deceive is a claim to his mental content, and the lawful character of the input will not cleanse the output of its testimonial character. To allow otherwise would permit the State to reconstruct the accused's entire mind and offer it in evidence if it starts from properly collected material, an abuse of process that *Selvi* was decided to prohibit [6].

4. The Algorithm as Witness

In the law of evidence, the witness makes a claim about a fact before the court, and he must be available for cross-examination with regard to it [19]. The system described in the Introduction is the former and not the latter: it occupies the witness's position while taking no oath, admitting no cross-examination of a meaningful kind, and, in the cases that matter most, recreating the content of the accused person's mind without his presence. That is why the existing doctrinal categories are ill-fitted to govern it. Scholars elsewhere have begun to ask whether machine conveyances should attract the testimonial safeguards the law reserves for human witnesses, most prominently Andrea Roth, whose account of machine testimony works within the confrontation and hearsay tradition of American evidence law [20]; the argument advanced here takes a different doctrinal route, locating the difficulty not in the right to confront an accuser but in the privilege against self-incrimination under Article 20(3), and in the distinctively Indian protection of the mind that *Selvi* secures.

One might want to compare the algorithm as witness with the expert witness, but the parallel does not hold at the crucial point. An expert gives the opinion that is based on a stated methodology, and that opinion, as well as methodology, can be challenged by the court; the court is entitled to probe the reasoning, the assumptions, and the candor of the witness [21,22]. The algorithm as witness provides the conclusion that is not amenable to such a probing, and there is no witness of any sort whose competence and honesty can be challenged, only a model whose behavior can be simulated in the post-analysis setting [15]. Nor is the difficulty cured by the human sponsor through whom the output is tendered. The State may reply that the output comes before the court through a human analyst, and that analyst can be cross-examined. But the analyst can only speak to the system's design, its validation, and its error rates in general. He did not make the inference, and he cannot answer for it. Examining him tests the sponsor's candor. It does not test the basis of the assertion [20]. The expert is accountable for his opinion; the algorithm is answerable to no one, and the law has never accepted a witness who cannot answer for what he asserts.

Neither is it forensic software under a new name. The forensic software is used to retrieve or measure some characteristic and report a match, a value, or a record that had already been existing. The algorithm as witness, on the contrary, generates the output: it produces a computed projection about the accused person, which did not exist before the system's work in any retrievable way [7]. The distinction between the retrieval and the generation is the distinction between an instrument and an author, and it is the latter who must take the witness's position.

The constitutional implication is even more ominous than the parallel with *Selvi* suggests at first sight. In *Selvi* the Court banned the use of narco-analysis and brain-mapping

because they invaded the mind's interior by overcoming its owner's will [6]. Machine inference invades this interior too, but it does not overcome the will but rather bypasses it completely: it requires neither the accused person's speech nor the physical activity of his body but the data that he had left during his lifetime. The privilege has always been there to protect the accused person from becoming the involuntary source of the information about his mind [23,24]. The inference makes him the source of that information without any compulsion, interrogation, or even the knowledge of the fact.

None of this is a condemnation of all algorithmic inference. A system that matches a fingerprint or identifies a face on the ground of physical characteristics creates the digital version of physical evidence, very close to that allowed in the *Kathi Kalu Oghad* case and the voice sample held admissible in *Ritesh Sinha vs. State of Uttar Pradesh* [25]. The dividing line lies between the inference about external characteristics, which do not come within the scope of the privilege, and the inference about the content of the mind. What must be avoided is the common practice of regarding the output of the machine merely as documentary evidence and hence neutral in character, since that would blur the line in the very place it needs to be kept clear.

Under Indian Law, the entryway into that error lies in the law of electronic evidence. The algorithmic output is entered into court proceedings as an electronic record, admissible under Section 65B of the Indian Evidence Act through certification and Section 63 of the Bharatiya Sakshya Adhiniyam [26–28], and the need for certification is obligatory [29,30]. What the certificate does is authenticate the record: it speaks to the normal functioning of the machinery behind the output, not to the truth of what the output asserts. A provision intended to ensure that a print-out reliably represents the computer's output cannot settle whether what the output says about the state of a person's mind is trustworthy.

5. The Problem in Comparative Perspective

The Artificial Intelligence Act adopted by the European Union represents the closest analogue to the regulation of the problem discussed above. The Act classifies the systems used in the law enforcement, including the individual risk assessment and certain types of biometric identification, as high-risk [31] (Annex III) and imposes on them the obligation of transparency, human supervision, and documentation that would render their outputs challengeable [32]. The Artificial Intelligence Act never mentions the privilege against self-incrimination. However, its requirement that a person shall be able to understand and challenge a significant automated decision relies on the same presumption as the proposition advanced in this paper from the other angle, namely, that the machine's inference about a person shall remain open to the challenge and may not be considered a fact just because the system made it.

The constitutional problem is similar to the Indian one under the different text in the United States. The American scholarly discussion of the confrontation of machine evidence addresses the question whether the output of a complex system can be cross-examined at all and whether due process demands certain procedures when a defendant is convicted based on a model whose workings are sealed [20,33,34]. No settled answer has yet emerged from the courts. However, what the literature makes clear is that the unconfrontable algorithmic accuser is not a unique feature of Indian doctrine but a structural element of the criminal procedure in general once the inference is introduced into the courtroom.

This would be closest to the Framework Convention on Artificial Intelligence of the Council of Europe, signed in Vilnius in September 2024, which is the first binding international treaty addressing this question [35]. It establishes that when a decision is adopted on the basis of or substantially through the use of an AI system, the person concerned has to be given sufficient information to be able to challenge the decision, and such procedural

safeguards are to apply in the case where the use of the system affects the individual's fundamental rights in a significant manner. Although the wording differs from Article 20(3), the Convention points in a similar direction on one point: an inference made using an algorithm about a person gives rise to a right to challenge it. The further step, that where the challenge is rendered impossible owing to the intrusion into the interior of the mind the answer is exclusion rather than disclosure, is the argument of this paper and not the text of the Convention. The fact that India is not a party to these instruments, and is not bound by them, is less important than the lesson that is to be learned from the comparative analysis: the Indian constitution, with its developed protection of the mind and the principle of proportionality, can offer a stronger solution to this problem.

6. Towards an Inference-Sensitive Reading

The test proposed is the two-part test, and the first part merely asks what is asserted in the inference. Where an AI system makes an assertion regarding the state of mind of the accused individual, his intentions, beliefs, knowledge, motives, or disposition, and the evidence used in making the assertion was not voluntarily obtained from the accused individual, then the assertion is testimonial in nature and is prohibited under Article 20(3) from being used against the accused individual [6]. This prohibition is unequivocal. Whether the inference made is correct or whether the crime is grave does not matter, since the protection of the accused is an absolute one and cannot be qualified by the extent of the State's interest.

The "without voluntariness" clause must be defined, since it is crucial in deciding cases. It does not refer to the lawfulness of the collection, nor does it refer to consent on the part of the accused for the collecting of data for another purpose. Data collected through a warrant is lawful collection but not voluntary production; data obtained from the record of a telecom operator or a platform have never been voluntarily produced by the accused individual; and data publicly posted are freely expressed, not voluntarily provided as testimony of mental interior. What the clause requires is the same thing that Selvi required, when it allowed narco-analysis conducted with consent and safeguards: a knowing and voluntary waiver with respect to the specific incriminating use, not just a general consent for the collecting of data followed by a waiver [6]. And the objection in the opposite direction is also answered: voluntariness of production does not render the machine's inference admissible, since consent to collecting is not consent to analysis. Two clarifications complete the first stage. The bar operates on use in evidence against the accused; it does not prevent investigators from following an inference as a lead, though investigative use remains subject to the discipline of Article 21 described below. And the stage governs claims about past mental states, the intent, knowledge, or belief accompanying the offence charged; a forward-looking risk score asserts propensity, and propensity is excluded, where it is excluded, by the ordinary law of evidence and relevance rather than by the privilege.

When the output asserts something else, an external fact, the identity of a face, the location of a device, the authorship of the physical mark, the privilege drops out and the second stage begins. Here the controlling precedent is *Justice K.S. Puttaswamy vs. Union of India*, which established the informational privacy as an integral part of Article 21 and proportionality as the standard of intrusion [36]. The court must determine whether the use of the system pursues a legitimate purpose, is reasonably connected to it, is necessary in the sense that no less intrusive method would serve the purpose, and is proportional to the privacy interest at stake, the four-point test of *Modern Dental College vs. State of Madhya Pradesh* [37]. The order of the two stages is not a matter of convenience: arriving at the proportionality before addressing the question of the character of the inference would

permit a sufficiently grave case to dissolve the categorical protection of the mind, which is the exact result that the sequence of the two steps is designed to avoid.

The second stage also involves an additional requirement which is not required by the first one: the output of the system must be understandable enough to be challenged by the defense. A conclusion whose basis is totally sealed cannot be admitted even if the privilege is not engaged, since the evidence which cannot be challenged cannot be tested, and the untested evidence about a person is not evidence but a mere assertion [16].

This is where the proposal diverges from the rules of disclosure. The main principle of almost all the AI regulation is to make the system transparent by disclosure, by more documents, more explanations, more access to the model. The inference about the mind requires the different treatment, since the objection is not the inability to see the mechanism of drawing the conclusion but the impermissibility of the conclusion itself. An inference prohibited by Article 20(3) cannot be cured by disclosure, and its opacity is immaterial once the privilege applies. The framework provided here excludes the most intrusive uses and disciplines the rest through the principles of proportionality and challengeability, reproducing in the inner structure of Indian constitutional law the right-to-challenge principle that emerges in the European Union AI Act and the Council of Europe's Convention [31,35]. The comparison also yields one concrete mechanism worth taking: the record-keeping and logging duties the European instrument attaches to high-risk systems are what make a right to challenge practical, since a conclusion can be contested only if a record of how it was produced exists [31] (Art. 12). India does not have to borrow the principle itself. It can derive a stronger one from Article 20(3) and Article 21, provided that it rejects treating machine inference as documentary evidence.

7. Conclusions

The emergence of AI-generated evidence introduces a new computational pathway through which the State may obtain assertions about an accused person's mind. The constitutional privilege against self-incrimination was developed to prevent the State from making the accused an involuntary source of evidence concerning his or her own mental content. Historically, this concern was principally associated with compelled speech. AI-generated inference creates a different technological pathway: the system can reconstruct aspects of mental content from digital traces without requiring the accused to speak, answer questions, or knowingly provide the alleged information. The resulting inference therefore challenges the traditional distinction between testimonial and physical evidence.

This paper has argued that the constitutional analysis should focus not only on how the underlying data were obtained, but also on what the AI system generates from those data. An external dataset may be lawfully collected and independently existing, yet the computational output derived from it may constitute an assertion about intention, knowledge, belief, motive, or another aspect of mental content. Treating the constitutional character of the output as identical to that of its inputs would allow computational systems to reconstruct the contents of an individual's mind through data that were never voluntarily provided as testimony.

The proposed inference-sensitive two-stage framework provides a structured response. At the first stage, the court should determine whether the AI-generated output asserts something about the accused's mental state and whether the underlying information was voluntarily provided by the accused for that specific incriminating use. Where the system generates an assertion concerning mental content without such voluntary testimonial participation, the paper argues that Article 20(3) prohibits its use against the accused [6]. The constitutional protection therefore attaches to the nature of the computational assertion,

rather than simply to the physical or digital character of the information from which it was generated.

Where the AI-generated output instead concerns an external fact—such as identity, location, or another physical characteristic—the Article 20(3) privilege does not operate in the same manner. The analysis then proceeds under Article 21 and requires proportionality, including legitimate purpose, rational connection, necessity, and proportionality to the privacy interest affected [36,37]. The framework further requires the output to be sufficiently understandable and documented to permit meaningful challenge. An AI-generated conclusion that cannot be examined, tested, or contested should not acquire evidentiary authority merely because it was produced by a computational system.

This requirement of challengeability is particularly important because AI systems can generate highly convincing outputs while remaining difficult to interpret or independently reproduce. Explainability mechanisms may identify relationships between inputs and outputs, but they do not necessarily establish that an inference is correct [15,16]. Similarly, automation bias can encourage decision-makers to treat computational outputs as more authoritative than conflicting human evidence [13,14]. In the criminal-justice context, these characteristics create a risk that algorithmic inference will acquire the appearance of objective fact without the procedural safeguards traditionally associated with testimonial evidence.

The framework therefore distinguishes between algorithmic inference as a computational process and algorithmic output as evidence. A machine-generated conclusion should not be treated as neutral documentary evidence simply because it is recorded electronically. Nor should the presence of a human analyst automatically cure the problem, because the analyst may be able to explain system operation and general performance without being able to testify to the actual inferential basis of the particular output. The evidentiary question must remain focused on the computational process that generated the assertion.

The broader significance of this analysis extends beyond Article 20(3). As AI systems become increasingly capable of inferring behavioural characteristics, intentions, preferences, and other latent attributes from heterogeneous digital traces, constitutional and evidentiary law will increasingly encounter outputs that did not exist in the underlying data in any directly observable form. The legal system must therefore distinguish between evidence that records a fact and computation that generates an assertion about a fact.

Ultimately, the question raised by AI-generated evidence is no longer only whether the accused can be compelled to testify. It is whether a computational system can become an algorithmic witness to the contents of the accused's mind when those contents were never voluntarily disclosed. The proposed framework provides a basis for addressing that question while preserving the constitutional distinction between protected mental content and external factual evidence, and while ensuring that computational authority does not substitute for evidence that can be meaningfully examined and challenged.

Author Contributions: Conceptualization, M.K. and S.S.P.R.; investigation, M.K.; writing-original draft preparation, M.K.; writing-review and editing, M.K. and S.S.P.R.; supervision, M.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: During the preparation of this manuscript, the author used Grammarly (Grammarly Inc., San Francisco, CA, USA; Premium version, accessed in 2026), QuillBot (QuillBot, Chicago,

IL, USA; Premium version, accessed in 2026) and Claude (Anthropic, San Francisco, CA, USA; Claude Sonnet 5, accessed in August 2026) for the purposes of checking grammar, spelling and punctuation, limited paraphrasing to improve readability, and obtaining critical feedback on the draft in the manner of a peer reviewer, respectively. The author has reviewed and edited the output and takes full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ranaware, V.; Mishra, R. Unquestioned Use of AI-Based Facial Recognition Technology in Criminal Investigations: Delhi Riots Lessons on Rights and Reliability. *Eng. Proc.* **2026**, *143*, 17. [CrossRef]
2. Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks. Available online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed on 31 August 2026).
3. Ferguson, A.G. *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*; New York University Press: New York, NY, USA, 2017.
4. Government of India. *Article 20(3)*; The Constitution of India; Ministry of Law and Justice: New Delhi, India, 1950.
5. State of Bombay vs. Kathi Kalu Oghad, AIR 1961 SC 1808 (Supreme Court of India). Available online: https://indiankanoon.org/doc/1626264/ (accessed on 25 August 2026).
6. Selvi vs. State of Karnataka, (2010) 7 SCC 263 (Supreme Court of India). Available online: https://indiankanoon.org/doc/338008/ (accessed on 28 August 2026).
7. Barocas, S.; Selbst, A.D. Big Data's Disparate Impact. *Calif. Law Rev.* **2016**, *104*, 671–732.
8. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E. On the Opportunities and Risks of Foundation Models. *arXiv* **2021**, arXiv:2108.07258.
9. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), Virtual Event, 3–10 March 2021; pp. 610–623.
10. Pasquale, F. *The Black Box Society: The Secret Algorithms That Control Money and Information*; Harvard University Press: Cambridge, MA, USA, 2015.
11. Kroll, J.A.; Huey, J.; Barocas, S.; Felten, E.W.; Reidenberg, J.R.; Robinson, D.G.; Yu, H. Accountable Algorithms. *Univ. Pa. Law Rev.* **2017**, *165*, 633–705.
12. Burrell, J. How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms. *Big Data Soc.* **2016**, *3*, 2053951715622512. [CrossRef]
13. Skitka, L.J.; Mosier, K.L.; Burdick, M. Does Automation Bias Decision-Making? *Int. J. Hum.-Comput. Stud.* **1999**, *51*, 991–1006. [CrossRef]
14. Parasuraman, R.; Manzey, D.H. Complacency and Bias in Human Use of Automation: An Attentional Integration. *Hum. Factors* **2010**, *52*, 381–410. [CrossRef] [PubMed]
15. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
16. Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef] [PubMed]
17. Citron, D.K.; Pasquale, F. The Scored Society: Due Process for Automated Predictions. *Wash. Law Rev.* **2014**, *89*, 1–33.
18. M.P. Sharma vs. Satish Chandra, AIR 1954 SC 300 (Supreme Court of India). Available online: https://indiankanoon.org/doc/1306519/ (accessed on 28 August 2026).
19. The Bharatiya Sakshya Adhiniyam, 2023, Section 142. Available online: https://indiankanoon.org/doc/76630180/ (accessed on 28 August 2026).
20. Roth, A. Machine Testimony. *Yale Law J.* **2017**, *126*, 1972–2053.
21. The Bharatiya Sakshya Adhiniyam, 2023, Section 39. Available online: https://indiankanoon.org/doc/32070916/ (accessed on 28 August 2026).
22. State of Himachal Pradesh vs. Jai Lal, (1999) 7 SCC 280 (Supreme Court of India). Available online: https://www.legalstreet.in/ratio/case/1999-insc-393/ (accessed on 28 August 2026).
23. Farahany, N.A. Incriminating Thoughts. *Stan. Law Rev.* **2012**, *64*, 351–408.
24. Ienca, M.; Andorno, R. Towards New Human Rights in the Age of Neuroscience and Neurotechnology. *Life Sci. Soc. Policy* **2017**, *13*, 5. [CrossRef] [PubMed]

25. Ritesh Sinha vs. State of Uttar Pradesh, (2019) 8 SCC 1 (Supreme Court of India). Available online: https://indiankanoon.org/doc/18061439/ (accessed on 26 August 2026).
26. The Indian Evidence Act, 1872, Section 65B (Repealed). Available online: https://indiacode.ecourtsindia.com/iea/section/65B/ (accessed on 27 August 2026).
27. The Bharatiya Sakshya Adhiniyam, 2023, Section 63. Available online: https://indiankanoon.org/doc/125020475/ (accessed on 28 August 2026).
28. Reddy, S.S.; Kaur, M. Silence After the Sanhita: Primary Evidence, Encrypted Data, and Article 20(3) under the Bharatiya Sakshya Adhiniyam, 2023. *Indian J. Law Leg. Res.* **2026**, *VIII*, 5712–5726. Available online: https://www.ijllr.com/post/silence-after-the-sanhita-primary-evidence-encrypted-data-and-article-20-3-under-the-bharatiya-s (accessed on 28 August 2026).
29. Anvar P.V. vs. P.K. Basheer, (2014) 10 SCC 473 (Supreme Court of India). Available online: https://indiankanoon.org/doc/187283766/ (accessed on 28 August 2026).
30. Arjun Panditrao Khotkar vs. Kailash Kushanrao Gorantyal, (2020) 7 SCC 1 (Supreme Court of India). Available online: https://indiankanoon.org/doc/172105947/ (accessed on 27 August 2026).
31. European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). *Off. J. Eur. Union* **2024**, 2024/1689.
32. Veale, M.; Zuiderveen Borgesius, F. Demystifying the Draft EU Artificial Intelligence Act. *Comput. Law Rev. Int.* **2021**, *22*, 97–112. [CrossRef]
33. Wexler, R. Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System. *Stan. Law Rev.* **2018**, *70*, 1343–1429.
34. State vs. Loomis, 881 N.W.2d 749 (Wis. 2016) (Supreme Court of Wisconsin). Available online: https://leagle.com/decision/inwico20160713i48 (accessed on 28 August 2026).
35. Council of Europe. *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*; CETS No. 225; Council of Europe: Strasbourg, France, 2024.
36. Justice K.S. Puttaswamy vs. Union of India, (2017) 10 SCC 1 (Supreme Court of India). Available online: https://indiankanoon.org/doc/91938676/ (accessed on 26 August 2026).
37. Modern Dental College and Research Centre vs. State of Madhya Pradesh, (2016) 7 SCC 353 (Supreme Court of India). Available online: https://indiankanoon.org/doc/49945774/ (accessed on 28 August 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.