Computational Humanities Research

www.cambridge.org/chr

Misaligning stories: Narrative unreliability, double noise, and feedback ethics in clinical AI

Rosa E. Martín-Peña

CAIMed – Lower Saxony Center for Artificial Intelligence and Causal Methods in Medicine & CELLS – Centre for Ethics and Law in the Life Sciences, Leibniz University Hannover, Germany.

Research Article

Cite this article: Martín-Peña Rosa E.. 2026. "Misaligning stories: Narrative unreliability, double noise, and feedback ethics in clinical AI." *Computational Humanities Research*, **2**:e14, https://doi.org/10.1017/chr.2026.10034

Received: 30 September 2025 UTC
Revised: 19 March 2026 UTC
Accepted: 20 May 2026 UTC

Keywords:
clinical AI; computational narratology; double noise; epistemic noise; feedback ethics; narrative unreliability

Corresponding author:
Rosa E. Martín-Peña;
Email: rosa-esther.martin-pena@cells.uni-hannover.de

Abstract

A worker with cancer is dismissed after algorithmic dashboards misread her treatment fatigue as cognitive decline. A stroke prediction model achieves 95 percent accuracy while missing every single stroke case. These are not edge cases – they are structural failures, produced when algorithmic distortions intersect with the cognitive variability of human judgment in the absence of feedback. This article calls that convergence double noise. Drawing on Shannon–Weaver's communication model, Kahneman and colleagues' concept of noise, and narratological theory – unreliable narration, focalization, paralepsis, emplotment, and situatedness – the article argues that predictive failures are not only statistical but narrative: computational systems operating under epistemic constraint produce false stories that resist correction. In both cases, the absence of meaningful feedback channels turns local distortions into entrenched misjudgments: the first through algorithmic dashboards that freeze discontinuous behavioral signals into an irrevocable story of cognitive decline; the second through routine design decisions – class balancing, metric selection, and threshold setting – that normalize the erasure of clinically decisive false negatives. By integrating narratological theory with computational methods and epistemic critique, the article positions double noise as a central challenge for clinical AI and advances feedback ethics as a normative orientation, calling for systems that preserve ambiguity, enable contestation, and institutionalize shared judgment in high-risk environments.

Plain language summary

When a cancer patient is dismissed because an algorithm misread her fatigue as cognitive decline, the problem is not only technical – it is communicative. This article examines how AI systems in medicine and work can misinterpret human behavior, and why feedback is the missing safeguard.

This article introduces the idea of double noise: when human judgment, which is variable, and algorithmic judgment, which is often simplified, overlap and reinforce each other. The result can be unfair or unsafe outcomes. Drawing on narrative theory, the article also shows how these systems do not merely misread people – they actively construct false stories by imposing coherence on incomplete signals, stories that then resist correction because no feedback channel exists to revise them.

The analysis draws on two case studies. The first is a fictional story about Leo, a worker whose illness led to short-term changes in her digital activity. Corporate dashboards misread these temporary signals as lasting decline, and her employer dismissed her without asking for context. Though fictional, this scenario is increasingly plausible in workplaces governed by algorithmic monitoring. The second examines a synthetic dataset on stroke prediction, designed for teaching. It illustrates how routine design choices – such as class balancing or metric selection – can obscure critical errors, especially false negatives.

Both cases show how, without feedback, small distortions harden into entrenched misjudgments. The article argues for systems designed to keep interpretive channels open: where professionals can question algorithmic outputs and individuals can challenge decisions that affect their health, work, or dignity. The framework developed here is intended as a transferable analytical tool – applicable wherever AI systems configure heterogeneous signals into consequential stories, suppressing the feedback that would allow those stories to be told differently.

Introduction

Uncertainty in clinical decision-making – whether among human professionals or in human–machine interactions – shapes not only patient outcomes but also the responsibilities and trust structures of healthcare institutions. In AI-based predictive models, this uncertainty is compounded by the opacity of black-box systems, which obscure class imbalance, unstable results, or sensitivity to data perturbations. Systematic reviews show a persistent gap between

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

claimed performance and clinical reliability (Char, Shah, and Magnus 2018; Rajkomar, Dean, and Kohane 2019; Shortliffe and Sepúlveda 2018; Topol 2019).

Human judgment is also inconsistent. As Kahneman, Sibony, and Sunstein (2021) demonstrate, even when presented with identical evidence, experts often reach different conclusions, influenced by contextual, emotional, and cognitive factors. This variability, called system noise, represents a form of epistemic uncertainty that persists independently of bias. Here, this variability is reframed as epistemic noise to distinguish it from algorithmic distortions. Computational models, on the other hand, introduce their own distortions, described here as algorithmic noise: errors derived from unbalanced datasets, preprocessing choices, or unstable predictions. When epistemic and algorithmic noise converge, distortions accumulate rather than cancel each other out, producing what can be called double noise.

The concept of noise has a dual heritage. In information theory, Shannon and Weaver (1949) defined noise as any disturbance that interferes with the transmission of a signal. In cognitive psychology, Kahneman, Sibony, and Sunstein (2021) demonstrated that noise in human judgment is stochastic, often invisible, and as harmful as bias. Together, these perspectives reveal how uncertainty permeates both technical and cognitive channels. In medicine, this manifests itself in inconsistent diagnoses, fluctuating triage decisions, or predictive blind spots – structural failures that threaten epistemic integrity (Kelly et al. 2019; Obermeyer and Emanuel 2016).

These failures are not merely technical, but also narrative. Human cognition imposes causal coherence on uncertain events, seeking closure even when it distorts reality (Nelson 1993). AI systems amplify this tendency: trained on decontextualized datasets and evaluated using aggregate metrics, such as accuracy or F-scores, they produce blind spots that remain hidden but carry high risk (Wiens et al. 2019; Wynants et al. 2020). Metrics, treated as objective, foster misplaced trust and epistemic deference, marginalizing alternative voices (Doshi-Velez and Kim 2017; Fricker 2007; Greene, Hoffmann, and Stark 2019). This article analyzes these narrative failures through three narratological frameworks that are not supplementary to the ethical argument but constitutive of it: *emplotment* (Ricoeur 1984), which reveals how algorithmic systems freeze discontinuous signals into irrevocable causal stories that foreclose renarration; *unreliable narration* and *focalization* (Booth 1961; Genette 1980), which expose how systems masquerade as omniscient narrators while operating from severely restricted epistemic positions; and *situatedness* (Herman 2009), which shows how the absence of feedback suppresses the discourse context that would allow those stories to be told differently. Understanding why predictive systems fail requires understanding how they narrate.

These epistemic failures do not arise in a cultural vacuum. From Orwell's *1984* (Orwell 1949) to *The Truman Show* (Weir 1998), narratives have long warned about systems that claim to know the subject better than they know themselves. Contemporary infrastructures, from wearable devices to clinical dashboards, inherit this imaginary: they anticipate, classify, and intervene, extending the logic of the panopticon (Foucault 1977). Prediction replaces consent, incorporating algorithmic governance into everyday life.

These dynamics are explored through two case studies. The first is a fictional narrative about Leo, a worker whose short-term changes in digital performance during illness are misinterpreted by workplace dashboards as a lasting decline, leading to her dismissal. The second examines a synthetic dataset on stroke prediction, in which design decisions obscure critical false negatives. Together, they demonstrate how, in the absence of feedback, double noise transforms uncertainty into entrenched misjudgment. To address this, the article draws on the Shannon–Weaver model of communication to trace how noise propagates across human and algorithmic channels, and advances feedback as both a technical and normative solution – systems designed to reopen the communicative loop, enable contestation, and support shared judgment in clinical AI (Barnlund 1970; Schramm 1954).

Leo's story: Narrative unreliability, double noise, and feedback loss

Although Leo is a fictional character, her situation reflects a plausible convergence between contemporary workplace surveillance and algorithmic misinterpretation, giving rise to dilemmas that need to be addressed before it is too late. Leo is in her early thirties and, to others, appears to be enjoying professional success, as suggested by her social media accounts and routines. She lives in a context where digital infrastructures permeate all aspects of everyday life. Yet she hides a secret: for a month, she has been battling a rare and aggressive form of breast cancer. Wanting to preserve her privacy and avoid disruption, she decides not to disclose her illness.

Corporate analytics tools, however, record many of her daily activities, continuously assessing her performance: keystrokes, login and logout times, breaks, email response rates, participation in online meetings, and even tone of voice and facial expressions during videoconferences. On days she worked from home, her activity patterns shifted more abruptly. She logged in later than usual, took more frequent breaks, and seemed less responsive in virtual meetings. These deviations lasted only two weeks, coinciding with the most acute phase of her treatment. Yet the automated dashboards continued to aggregate all these signals, constructing a coherent narrative of "declining cognitive performance."

In other words, metrics designed to capture productivity were reinterpreted as evidence of mental instability or cognitive decline. This framing was entirely misleading: Leo's actual condition was a localized breast cancer, a serious but somatic illness that did not impair her cognitive capacities. Fatigue and temporary irregularities in her schedule reflected the embodied toll of treatment, not a loss of reliability or judgment. Nevertheless, the system translated bodily suffering into a false story of psychological deficiency, a misclassification that her employer accepted without question.

This translation of bodily suffering into a false story of psychological deficiency is not merely an ethical failure – it is, in precise narratological terms, a failure of narration. The corporate dashboard functions as an *unreliable narrator* in the sense defined by Booth (1961)[1]: an institutional voice whose account of events is systematically distorted by incomplete information and structural bias.

[1] Booth's original concept presupposes a gap between the narrator's account and the implied author's norms, often mediated by irony or moral distortion. The corporate dashboard, however, instantiates a distinct variant: not a narrator who distorts knowingly, but one whose epistemic incapacity prevents it from recognizing its own distortion. This aligns more closely with what subsequent narratology has theorized as *cognitively restricted* narration – unreliability produced not by duplicity but by structural ignorance (Nünning 1999). The distinction matters: ironic unreliability implies an implied author who "knows better"; epistemic unreliability implies a system that cannot.

https://doi.org/10.1017/chr.2026.10034 Published online by Cambridge University Press

Leo's illness, invisible to the data infrastructure, constitutes precisely the kind of knowledge gap that defines narrative unreliability. This unreliability is compounded by a problem of *focalization* (Genette 1980, 189–90): the dashboard operates under the illusion of zero focalization – an ostensibly omniscient perspective that claims to capture Leo's full cognitive and affective profile. Yet Leo herself holds the only epistemically decisive information: the embodied reality of her diagnosis. What appears as objective surveillance is in practice a severely restricted point of view masquerading as total knowledge. In Genette's terms, this constitutes a structural *paralepsis* (Genette 1980, 195–8): the system gives more information than its actual epistemic position authorizes, projecting a verdict of cognitive decline that exceeds what its restricted access to behavioral proxies could ever justify. The dashboard does not merely misread Leo – it oversteps the bounds of what any narrator operating from its focal position could legitimately claim to know.

This deference materializes concretely: her manager interpreted these signals as evidence of unreliability and even depressive disorder. Algorithmic management systems compress temporary fluctuations into lasting traits, transforming irregularities into supposedly objective truths about workers' abilities – a dynamic that exemplifies double noise in its most direct form. First, algorithmic noise: incomplete and decontextualized signals recombined into misleading patterns. Second, epistemic noise: the manager perceives a decline in productivity but never asks Leo for context, choosing instead to dismiss her without explanation. As Gandini (2019) argues, the absence of genuine feedback channels in digitalized workplaces shifts power asymmetrically: workers are judged by indicators that cannot be negotiated.

Leo's cancer – an embodied, context-rich reality – remains invisible to the data infrastructure. The systems monitoring her never captured medical evidence; instead, they reconstructed behavioral proxies into an image of emotional instability. This aligns with what Zuboff (2019) calls "surveillance capitalism": predictive logics that extract behavioral data not to understand individuals, but to anticipate and control their trajectories. In Leo's case, anticipatory modeling exerts a preemptive judgment of unreliability, bypassing consent and silencing her lived experience. The result is dismissal. Despite existing safeguards in Europe – the AI Act's ban on emotion recognition in the workplace (European Parliament and Council of the European Union (EU) 2024) and the GDPR's protection of sensitive data (European Parliament and Council of the European Union 2016) – her company circumvents them by relying on productivity and behavioral metrics. These indicators function as affective proxies without being defined as such, creating a loophole that enables covert medicalized classifications.

Leo's case reveals the dangers of narrative compression in AI-mediated management. Momentary fluctuations in her work patterns, lasting only two weeks, are reinterpreted as permanent deficits. As Ajunwa, Crawford, and Schultz (2017) show, automated surveillance systems rarely distinguish between short-term deviations and long-term trends, making them structurally biased against workers in vulnerable situations. Without feedback, distortions solidify into definitive judgments. Leo is abruptly dismissed without being able to raise her voice or say goodbye to her colleagues, in line with organizational routines that minimize negotiation (Bürgi-Tian and Boehm 2021; Sennett 1998).

These failures of compression are, in narratological terms, also failures of *situatedness* (Herman 2009): the principle that any narrative representation must be interpreted in light of a specific discourse context. As Piper, So, and Bamman (2021) formalize this concept for computational analysis, narrativity is not only a textual property but also a cognitive and contextual process – one that cannot be reduced to pattern recognition across decontextualized tokens. The corporate dashboard produces precisely such a desituated narrative: it processes Leo's behavioral signals as if they were context-free data points, stripping away the embodied, temporal, and cultural occasion that would be necessary for meaningful interpretation. The absence of feedback is, in narratological terms, the absence of situatedness – the suppression of the discourse context that would allow Leo's story to be told differently.

Taken together, these narratological failures – unreliable narration, false focalization, and narrative desituatedness – compound into double noise. Algorithmic noise produces misleading conclusions, and epistemic noise reinforces them through uncritical deference to metrics. In the absence of feedback, Leo's lived reality cannot re-enter the decision-making process. This raises a broader question of responsibility: is the misjudgment attributable to the system that compresses signals into misleading narratives, the manager who defers to them without context, or the cultural norms that penalize silence? Without feedback, this distribution of responsibility remains opaque, further deepening the injustice.

Emotion, misrecognition, and epistemic noise

In recent years, there has been a marked increase in attempts to quantify emotional states using biometric and behavioral markers. Many contemporary AI systems still rely on reductionist theories, most notably those influenced by Ekman's model of "basic emotions" (Ekman 1992; Ekman and Friesen 1969), which treat affect as a set of universal, biologically hardwired categories. Much of affective computing depends on corpora annotated with Ekman's categories – happiness, sadness, anger, fear, surprise, and disgust – despite decades of critique and evidence of cultural variability. Lisa Feldman Barrett offers one of the most influential counterarguments, insisting that there are no universal "fingerprints" of emotion. Instead, emotions are socially and culturally constructed categories that emerge from context-dependent interpretations of bodily signals (Barrett 2017). This "variety is the norm" poses a major challenge for algorithmic systems, which continue to treat emotions as static labels tied to physiological or behavioral patterns.

This conceptual mismatch is not only epistemological but also narratological. A system trained on Ekman's universal categories functions as an *unreliable narrator* in the precise sense theorized by Nünning (1999): not a narrator who distorts knowingly, but one whose epistemic architecture prevents it from recognizing its own distortion. It produces a story about the subject's affective state – one the subject herself would recognize as false – without any capacity to register the gap between its account and the lived reality it claims to represent. This is cognitively restricted narration in its most structural form: the model cannot know what it cannot see, and what it cannot see is precisely what Barrett shows to be constitutive of emotional meaning.

This unreliability is compounded by a problem of *focalization*. Ekman-based systems operate under the illusion of zero focalization (Genette 1980): they claim to capture the subject's full affective profile from facial micro-expressions, vocal modulation, or interaction rhythms, projecting a totalizing epistemic authority that their actual access to surface proxies could never justify. In Genette's terms, this constitutes a structural *paralepsis*: the system gives more interpretive weight to its output than its restricted focal

https://doi.org/10.1017/chr.2026.10034 Published online by Cambridge University Press

position authorizes. What appears as objective affect recognition is in practice a severely constrained point of view masquerading as total knowledge.

The result is also a failure of *situatedness* (Herman 2009). Barrett's central claim – that emotions are context-dependent interpretations of bodily signals – is, in narratological terms, a claim about situatedness: emotional meaning cannot be detached from the specific embodied, relational, and cultural occasion that produces it. A model trained on decontextualized corpora treats as context-free what is irreducibly situated, stripping away the discourse context that would be necessary for meaningful interpretation. When Leo's fatigue is read as disengagement, or her silence as unreliability, the system is not merely misclassifying – it is producing a desituated narrative that forecloses the very context that would make her signals legible.

Finally, and most consequentially, this is a failure of *emplotment* (Ricoeur 1984). Where narrative configuration should synthesize heterogeneous affective signals into a meaningful and revisable whole, Ekman-based systems freeze discontinuous and culturally variable expressions into single, irrevocable categorical verdicts. What emplotment, properly understood, keeps open – the possibility of renarrating, of integrating new context into a revised account – the algorithmic system forecloses. A momentary shift in vocal tone becomes "anger"; a brief absence of eye contact becomes "disengagement"; a two-week fluctuation in productivity becomes "cognitive decline." Each of these moves enacts the same narrative violence: the compression of a living, revisable signal into an irreversible story.

This conceptual mismatch produces epistemic noise: a systematic gap between the richness of lived affective experience and the reductive formats through which it is captured and classified. When algorithmic simplifications then meet uncritical institutional deference, the result is double noise: epistemic variability in human affect flattened into algorithmically rigid categories, and those categories cemented into definitive judgments without contestation. What is missing is not only accuracy but also feedback – the channel through which subjects could recontextualize how their emotional states are inferred, and through which errors could be re-situated in lived experience rather than hardening into institutionalized misrecognition.

The EU Artificial Intelligence Act (2024) (European Parliament and Council of the European Union (EU) 2024) explicitly prohibits the use of emotion recognition in workplaces and education. Yet this safeguard addresses only direct recognition, not the indirect reconstruction of affective states through behavioral proxies. Typing speed, login times, geolocation traces, or email responsiveness can all serve as soft indicators of cognitive or affective states, enabling what McStay (2018) terms "emotional AI." In such cases, emotion is not formally detected but statistically inferred, and decisions are taken based on deviations from behavioral norms. The prohibition is thus respected in form but circumvented in practice – and the narratological failures of unreliable narration, false focalization, desituatedness, and foreclosed emplotment operate undisturbed beneath the regulatory threshold.

In Leo's case, this also amounted to cultural injustice. Coming from an Asian background where collective harmony is valued over individual disclosure, her silence was not simply a personal choice but a culturally situated communicative act – one that the system's predefined taxonomies rendered illegible. As Fricker (2007) argues, such dynamics constitute epistemic injustice: the systematic exclusion of certain voices from credibility. In more individualist contexts, she might have taken sick leave without penalty; her decision to keep working – shaped by collective norms of discretion – was reinterpreted as unreliability, sealing her exclusion. Feedback could have reopened the communicative loop; its absence transformed epistemic noise into institutionalized misrecognition.

Between prediction and explanation: Revisiting the "broken leg"

If anything has marked the history of research on human judgment, it has been a constant tension between statistical prediction and human intuition. In the 1950s, Meehl (1954) famously demonstrated that actuarial models consistently outperformed clinical experts in predicting outcomes, such as academic success or psychiatric relapse. Yet Meehl himself acknowledged a crucial limitation, illustrated by what became known as the "broken leg case." Suppose a man usually goes to the movies every Tuesday, and a statistical formula predicts his attendance. If we learn, however, that he has a broken leg, this singular, context-rich event – though statistically rare – completely invalidates the prediction. The lesson: even highly accurate models can fail spectacularly when faced with anomalies that resist probabilistic encoding.

In clinical AI, these anomalies are not hypothetical, but rather close to the norm. Patients with rare diseases, comorbidities, or atypical trajectories often fall outside the training distribution of predictive systems. These cases, which can be considered by analogy as "broken leg" cases, are frequently treated as outliers or noise and removed during preprocessing in the name of model optimization. Yet they may represent precisely those patients most in need of medical care. What appears to be a statistical inconvenience is, in fact, an ethical, social, and clinical failure: the tension between model generalization and context specificity resurfaces not as an epistemological footnote, but as a central challenge to justice in healthcare.

Recent work in machine learning has underscored similar concerns. Researchers in anomaly detection argue that discarding rare or unusual cases can distort both training and evaluation, especially in medical contexts, where anomalies may carry decisive clinical meaning. Others highlight how anomaly detection and outlier exposure techniques could improve robustness by explicitly training models to recognize and flag such cases rather than silently misclassifying them (Hendrycks, Mazeika, and Dietterich 2019; Machado, Silva, and Gama 2023). These approaches reveal a structural analog to the "broken leg case": outliers are not statistical curiosities but potential signals of high-risk conditions.

Leo's story embodies one such "broken leg" case. Her temporary changes in digital behavior, caused by the embodied toll of cancer treatment, fell entirely outside the expectations of workplace algorithms. The system recoded her suffering into a misleading narrative of cognitive decline, rendering her condition illegible within dominant predictive frameworks. This is precisely the kind of anomaly that cannot be anticipated through probabilistic reasoning but demands contextual interpretation. In Herman's terms, this constitutes a failure of situatedness: the system lacks access to the embodied and biographical discourse context that would make the anomaly legible. In Leo's case, the absence of feedback made such interpretation impossible: her lived signals were treated as noise, erased from the decision process.

This unresolved conflict sets the stage for the next case: a pedagogical analysis of stroke prediction using a widely circulated synthetic dataset. Unlike Leo's story, where the anomaly was biographical and context-specific, the stroke dataset illustrates how anomalies can be structural, produced not by patients themselves

https://doi.org/10.1017/chr.2026.10034 Published online by Cambridge University Press

but by modeling conventions. The dataset was originally designed for teaching programming and machine learning practices, not for clinical deployment. Its purpose was educational, yet the way it is typically used – in exercises on classification, clustering, and model evaluation – reveals how routine design choices, such as data balancing, metric selection, or threshold setting, can systematically obscure false negatives. What is sidelined for didactic neatness are precisely the patients who, in a real-world setting, would be at highest risk. In this sense, the stroke dataset functions as a "broken leg case" by design: its pedagogical simplifications intentionally strip away clinical nuance, creating blind spots that mirror real-world failures of predictive systems.

Crucially, the "broken leg case" underscores that predictive failures are not only statistical but also communicative. When anomalies – whether lived or structural – are treated as noise, the signal of a patient's condition is lost in translation. What fails is not simply the model's accuracy, but the chain through which information is encoded, transmitted, and decoded. Without a channel for feedback, these signals cannot re-enter the communicative loop; they are silenced before they can challenge the authority of the model.

The broken leg case is, in narratological terms, not only an epistemological failure but also a failure of *focalization*. The actuarial model operates under the same illusion of zero focalization that governs Leo's corporate dashboard: it projects a totalizing epistemic authority – the man will go to the cinema on Tuesday – from a severely restricted focal position that cannot accommodate what it cannot see. The broken leg is precisely the kind of embodied, context-rich signal that resists probabilistic encoding, and its invisibility to the model constitutes a structural *paralepsis*: the system gives more interpretive weight to its prediction than its actual epistemic access could ever authorize. This is also a failure of *emplotment* (Ricoeur 1984): the model freezes a dynamic, revisable behavioral pattern – Tuesday cinema attendance – into an irrevocable causal story that forecloses the very anomaly that would make it false. What Meehl's case reveals, then, is not merely a statistical limit but a narrative one: predictive systems fail not only when their data are incomplete, but also when their architecture cannot renarrate in the face of lived reality.

Leo's case is, in Meehl's terms, a "broken leg" case: a singular, context-rich anomaly that invalidates the prediction not because the model is technically deficient, but because its epistemic architecture – operating under the illusion of zero focalization – cannot accommodate what it cannot see. The absence of feedback is what prevents the broken leg from re-entering the communicative chain – and what allows a failed emplotment to solidify into verdict. The stroke dataset that follows enacts the same logic at a structural level: here, the anomaly is not biographical but methodological, produced not by a patient's hidden condition but by the modeling conventions themselves. In both cases, what disappears is not noise in the technical sense – it is the signal that matters most.

What is missing, then, in both Leo's dismissal and in the pedagogical treatment of stroke prediction, is a robust mechanism of feedback. Feedback would not erase uncertainty, but it would allow context to interrupt prediction, giving rare conditions and exceptional cases the possibility of recognition. This move from descriptive model to normative claim is not arbitrary: once communication is understood as the medium through which clinical judgment is formed, the conditions that distort or foreclose it become ethical stakes, not merely technical ones.

Shared judgment and the ethics of feedback

The key issue in algorithmic and cognitive noise is not merely one of accuracy, but of communication: as information travels through systems of judgment, some signals are amplified while others are deleted, ignored, displaced, or mistranslated. What disappears in this process is the patient's living signal, the embodied reality that resists neat encoding. Noise, in this sense, is not only technical but also cultural: signals may be dismissed because they do not align with normative expectations of productivity, compliance, or disclosure.

To grasp how such distortions arise and persist, this section turns to the Shannon–Weaver model of communication (Shannon and Weaver 1949). Originally conceived for telegraph lines and technical signal transmission, the model offers a surprisingly apt perspective for examining AI-supported decision-making. It reminds us that every prediction is also a message: something must be encoded, transmitted, and decoded, always under the threat of noise. Noise, in this sense, is not a minor glitch but a structural form of uncertainty that, if ignored, can open the way to epistemic injustice.

The model distinguishes six components – source, encoder, channel, noise, decoder, and destination – each a potential fracture point where meaning may shift. Later refinements by Schramm (1954) and Barnlund (1970) introduced feedback as the element that transforms linear transmission into dialogue. In clinical AI, feedback is not a decorative feature; it is the moment where signals are tested against lived knowledge, where patients and practitioners can resist, correct, or reinterpret algorithmic claims. Without such loops, distortions solidify into decisions, producing unfair norms that become increasingly difficult to contest. Feedback is therefore not only a safeguard against error but also a mechanism that clarifies responsibility, redistributing accountability across patients, practitioners, and systems.

Seen from this perspective, noise is not merely an error to be filtered out. When algorithmic variability and human bias converge without the corrective force of feedback, noise does not diminish – it doubles. This phenomenon of *double noise* explains why failures persist, why judgments misalign, and why epistemic injustice can become entrenched in clinical AI.

Figure 1 reinterprets the Shannon–Weaver model for clinical AI, illustrating how algorithmic and epistemic noise can propagate through each stage when feedback is absent or suppressed.

Mapping these distortions across the communicative chain reveals that noise does not accumulate randomly – it concentrates at the stages where human judgment and algorithmic output intersect without mediation. Table 1 illustrates this pattern, showing how algorithmic and cognitive noise converge at each stage when feedback is absent. Unlike other stages, feedback is transversal: it can retroactively alter interpretation across the entire chain.

Seen in clinical practice, each element of the Shannon–Weaver chain has a direct analog. The *information source* may be a patient history or a clinical trial dataset; the *encoder* a diagnostic test or algorithmic feature extractor; the *channel* an electronic health record system; and the *decoder* a clinician interpreting model outputs. At each stage, algorithmic and cognitive noise intertwine, with feedback loops determining whether errors are corrected or silently reinforced. This distributed framing shifts the question of accountability: failures do not belong to a single faulty model or a careless clinician, but to the fragile links between them. If the communicative chain is where meaning is formed – and where

https://doi.org/10.1017/chr.2026.10034 Published online by Cambridge University Press

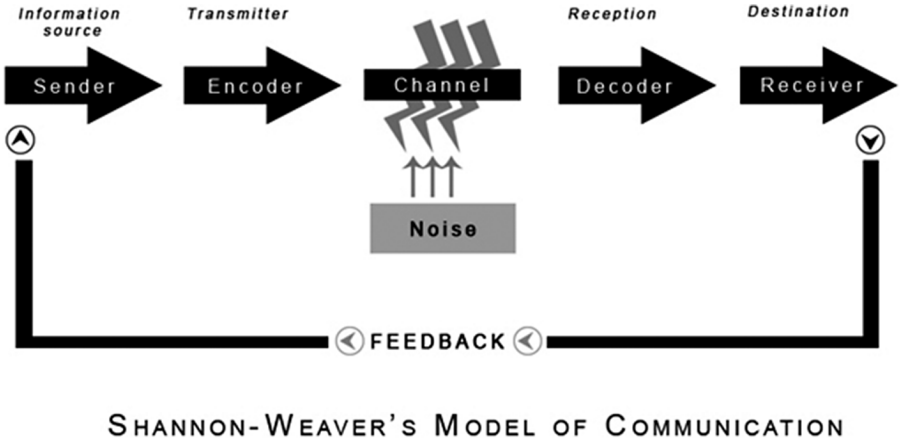


Figure 1. Reinterpretation of the Shannon–Weaver model in the context of clinical AI. Double noise propagates across multiple stages when feedback is weak or absent.

Table 1. Sources of algorithmic and cognitive noise across the Shannon–Weaver model in clinical AI

Stage	Algorithmic noise	Cognitive noise
Information source	Biased, incomplete, or non-representative training data	Selective attention, prior assumptions, and limited anamnesis
Encoder	Faulty sensors, data transformation errors, and preprocessing artifacts	Inconsistent documentation and ambiguous clinical notes
Channel	Transmission delays, interface mismatches, and system latency	Fatigue, cognitive overload, and environmental distractions
Noise source	Class imbalance, adversarial examples, and model variance	Stress, time pressure, implicit bias, and contextual misalignment
Decoder	Non-interpretable outputs and instability in predictions	Framing effects, misinterpretation of probabilities, and reliance on heuristics
Destination	Misclassification due to poor generalization and interface confusion	Overreliance on system outputs, mistrust, and interpretive shortcuts
Feedback	Neglected feedback signals, retraining misalignment, and audit failure	Suppression of dissent, epistemic deference, and lack of critical reflection

it can be lost – then the ethical imperative is not to optimize individual nodes but to protect the integrity of the chain as a whole. As Mittelstadt et al. (2016) and Ananny and Crawford (2018) emphasize, governance failures often arise not from the absence of data but from the breakdown of communication and accountability.

Clinical AI, therefore, must be designed for *shared judgment* – a mode of decision-making in which clinicians, patients, and algorithmic systems participate as co-interpreters rather than as isolated producers and consumers of scores. Unlike fixed decision thresholds, which cannot account for context-specific anomalies, shared judgment institutionalizes the conditions under which uncertainty can be named, negotiated, and revised. In narratological terms, it restores what Herman calls *situatedness* (Herman 2009) – the feedback channels through which a system's output can be reinterpreted in light of the specific discourse context that produced it.

This requires embedding iterative feedback that enables contestation and recalibration across the entire communicative chain. The ethic that follows is one of *feedback ethics*: systems that resist premature closure, foreground plurality, and institutionalize opportunities for negotiation (Doshi-Velez and Kim 2017; Fricker 2007). This normative move is not external to the Shannon–Weaver model but internal to it: once communication is understood as the medium through which clinical judgment is formed, the conditions that distort or foreclose it are not merely technical failures but ethical ones. The case study that follows makes this concrete.

With this theoretical framework established, the article now turns to a technical case study on stroke prediction. This case illustrates how communicative fractures identified in the Shannon–Weaver model materialize in practice: routine preprocessing steps and metric choices, once seen as neutral, propagate double noise and foreclose feedback.

Double noise and blind spots in stroke prediction

To complement this narrative, the present section turns to a technical case study: the analysis of a pedagogical dataset for stroke prediction. Although synthetic and simplified, this dataset cannot function as a direct proxy for real-world clinical deployment. Its value is strictly pedagogical: it demonstrates how structural blind spots emerge from routine modeling conventions, not that such blind spots occur with this frequency or severity in clinical practice. The generalization it supports is therefore methodological rather than epidemiological – it shows *how* erasure can be normalized by default pipelines, not *how often* this occurs in deployed systems.

https://doi.org/10.1017/chr.2026.10034 Published online by Cambridge University Press

This distinction matters: the ethical argument does not depend on the dataset's clinical representativeness, but on the structural logic it makes visible. Importantly, this dataset was never intended for clinical use but for teaching programming, with a special focus on false positives. Its educational intent highlights a paradox: even datasets created for pedagogical clarity can reproduce structural blind spots when feedback is absent. Tracing this logic requires applying the same narratological concepts introduced in the preceding sections – unreliable narration, focalization, paralepsis, and emplotment – to the pipeline itself: not as metaphors imported from literary theory, but as analytical tools that make visible the communicative failures that aggregate metrics routinely conceal.

Methodology and preprocessing decisions

These preprocessing and modeling steps echo long-standing challenges in the machine learning literature on class imbalance and evaluation (Chawla et al. 2002; He and Garcia 2009; Saito and Rehmsmeier 2015), but here they are revisited from the perspective of communicative breakdowns in clinical AI. By tracing how imbalance and default routines interact with interpretive blind spots, the case highlights how technical conventions become ethical stakes when applied to health contexts.

The dataset was divided into stratified training and test sets (70/30) to preserve the rare positive class. Missing values in continuous variables, such as BMI, were imputed with the median; categorical attributes were encoded using one-hot vectors; and numerical features were standardized to zero mean and unit variance for comparability across algorithms. No feature engineering was applied – an intentional choice to keep the pipeline minimal and transparent.

Each of these steps, although standard, introduced subtle distortions. Median imputation assumes that "average bodies" can represent missing ones, erasing heterogeneity. One-hot encoding breaks nuanced categories into rigid binaries. Standardization imposes a geometric comparability that privileges certain relations while marginalizing others. Even the choice of stratification, while protecting class proportions, fixes rarity into the dataset rather than questioning its interpretive implications. As He and Garcia (2009) emphasize, such preprocessing conventions interact problematically with class imbalance, reinforcing patterns of invisibility for minority outcomes. Popular remedies, such as SMOTE oversampling (Chawla et al. 2002), illustrate that technical fixes exist, but their absence in pedagogical datasets normalizes the erasure of rare cases from the start. Together, these moves exemplify *algorithmic noise*: distortions created when lived variability is forced into mathematically convenient formats.

At the same time, their routine acceptance generates *epistemic noise*. Defaults are often treated as methodological inevitabilities, masking the fact that different imputation strategies, encodings, or balancing methods would shift where errors fall. Crucially, in health contexts, such errors are not symmetric: false negatives carry radically different risks than false positives. As Saito and Rehmsmeier (2015) demonstrate, accuracy and ROC curves can be deeply misleading under imbalance, obscuring systematic failures of recognition that are visible only in precision–recall analysis. Preprocessing, therefore, is not a neutral preparation stage but the first site where double noise enters the modeling chain.

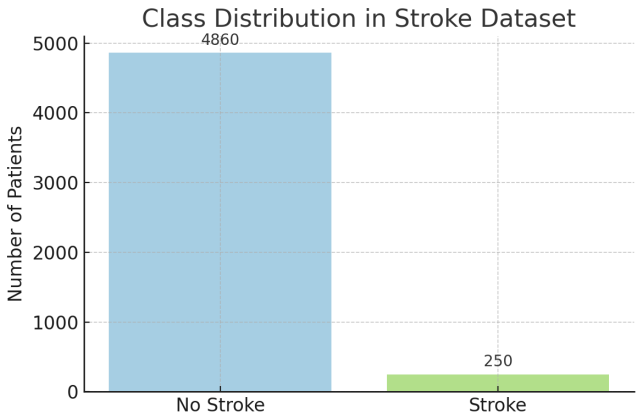


Figure 2. Class imbalance in the stroke dataset. The vast majority of patients are labeled as *No Stroke*, while only a small minority experienced a stroke. This imbalance inflates accuracy and conceals failures in detecting critical cases.

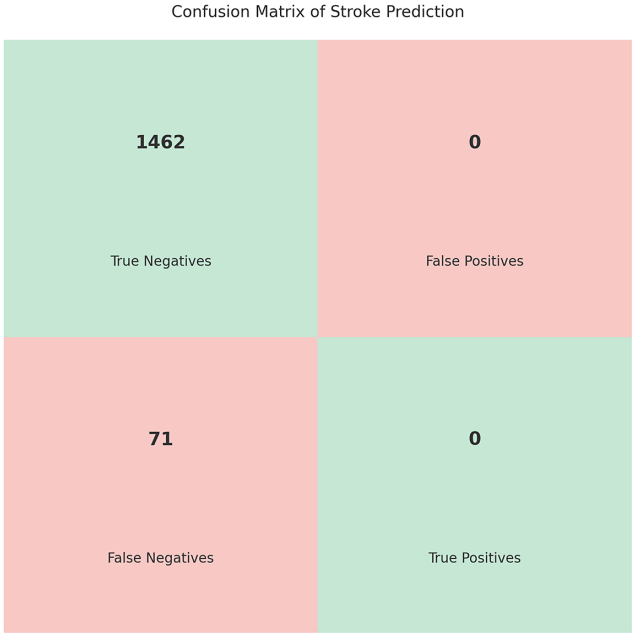


Figure 3. Confusion matrix of the Random Forest model. All stroke cases in the test set were misclassified as non-stroke, illustrating a total failure to detect the minority class.

Supervised models: Results and critical analysis

A Random Forest classifier was trained with default settings, except for a maximum tree depth of 3 to preserve interpretability and reduce overfitting. Stratified sampling maintained the original class distribution.

Overall performance appeared excellent – 95.4 percent accuracy and a weighted F_2 score of 94.5 percent. Yet the model failed to identify a single stroke case in the test set. All positive cases were misclassified as negative (Figures 2 and 3).

Performance summary:

- **Accuracy:** 95.4 percent.
- **Weighted precision:** 90.9 percent.
- **Weighted recall:** 95.4 percent.
- **F_2 score:** 94.5 percent.

This failure is not only statistical but also narratological. The Random Forest pipeline operates, like Leo's corporate dashboard,

https://doi.org/10.1017/chr.2026.10034 Published online by Cambridge University Press

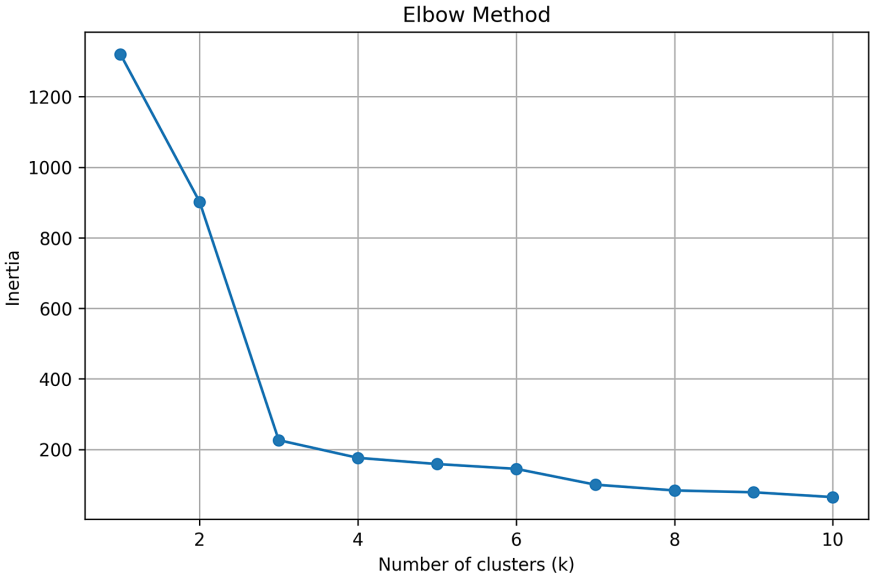


Figure 4. Elbow method applied to the stroke dataset. The inflection point around $k = 3$ suggests three latent clusters with distinct risk factor profiles.

under the illusion of zero focalization: it produces a verdict of "no stroke" with the same epistemic authority that the dashboard projected "cognitive decline," while both systems lack access to the very information that would make their outputs meaningful. In Genette's terms, this constitutes a structural *paralepsis* at the level of the model itself: the classifier gives more interpretive weight to its output than its restricted epistemic position – trained on imbalanced proxies, evaluated on aggregate metrics – could ever authorize. The pipeline is, in this sense, an *unreliable narrator*: it produces a story of clinical safety that the patients it misclassifies would recognize as false, not through duplicity but through structural ignorance of what it cannot see. The absence of feedback forecloses the possibility of renarration: once the model outputs "no stroke," no channel exists through which the false negative can re-enter the communicative chain as a signal worth interpreting.

These numbers conceal the clinical failure: the model "succeeds" by correctly classifying the majority class while systematically ignoring the minority. This behavior is well known in the literature on imbalanced learning, which shows that class weighting, oversampling, or threshold adjustment alone rarely resolve the deeper problem of underrepresented cases (Chawla et al. 2002; He and Garcia 2009). Even when these remedies improve numerical balance, they do not address the ethical implications of systematically missing rare but clinically decisive cases. To contextualize this failure, a counter-scenario was applied: a Logistic Regression model trained on SMOTE-balanced data. The results illustrate the trade-off that "normalize erasure" conceals. With SMOTE, recall for the minority class improved substantially – 42 of 71 stroke cases were correctly identified, compared to zero in the default pipeline. However, this gain came at significant cost: accuracy dropped from 95.4 percent to 68.3 percent, and false positives increased to 457. This contrast demonstrates that the claim of "normalization of erasure" is not metaphorical but empirically grounded: the default pipeline does not merely underperform – it systematically renders the minority class invisible while projecting an illusion of reliability.

Moreover, standard evaluation metrics can reinforce this illusion of success. As Saito and Rehmsmeier (2015) demonstrate, accuracy and ROC curves are misleading under severe class imbalance, whereas precision–recall analysis provides a more informative view of minority-class performance. In a clinical context, such blind spots could mean missed interventions, increased morbidity, or preventable deaths. As Hooker et al. (2021) argue, framing bias solely as a "data problem" ignores how evaluation pipelines themselves sustain systemic failures. Pedagogical datasets exacerbate this by teaching students that high accuracy is sufficient, reinforcing illusions of reliability. Similarly, Ghassemi, Oakden-Rayner, and Beam (2021) caution against the false hope that explainability or transparency alone will resolve these failures, stressing that deeper attention to evaluation practices and error distributions is needed.

Unsupervised clustering and latent profiles

To probe heterogeneity beyond binary classification, unsupervised clustering was applied using K-means on normalized features: `age`, `BMI`, and `avg_glucose_level`. After excluding rows with missing values, the Elbow Method indicated an optimal $k = 3$ (Figure 4).

Although these clusters did not align with stroke outcomes, they revealed subgroups with distinctive combinations of risk factors. One cluster grouped younger patients with elevated glucose levels, another older individuals with high BMI but no comorbidities. These profiles highlight that dichotomous classification can obscure clinically relevant differences. Unsupervised learning, while not predictive, offers a diagnostic lens: it surfaces ambiguities that supervised pipelines suppress. For pedagogy, this contrast is instructive – it shows that the same dataset can yield either silence (via classification) or plural signals (via clustering), depending on how feedback is conceptualized.

From technical blind spots to situated meaning

This case study demonstrates how predictive systems can fail silently. Despite high reported accuracy, the absence of any true positive predictions reveals a representational void: the patient disappears from the model's output, and the communicative loop

https://doi.org/10.1017/chr.2026.10034 Published online by Cambridge University Press

remains closed. Without feedback, distortions become entrenched as truths.

At the same time, unsupervised clustering surfaced latent patterns that binary classification erased. This divergence underscores the need to treat evaluation not as a matter of aggregate metrics but as part of a communicative chain. Accuracy or F-scores can project certainty while masking failures of recognition. The same logic of emplotment is at work here (Ricoeur 1984): the pipeline fixes heterogeneous clinical signals into a single irrevocable verdict, foreclosing the possibility of revision. In high-stakes contexts, such illusions of reliability carry severe consequences.

As Rajkomar, Dean, and Kohane (2019), Topol (2019), and Wiens et al. (2019) emphasize, technical validity must always be tested against clinical relevance, with particular scrutiny on error types that affect patient safety.

The lesson is twofold. First, preprocessing and modeling decisions, even in synthetic and pedagogical settings, enact ethical stakes: they distribute visibility and invisibility across patients. Second, evaluation itself must be reframed: predictions are not isolated outputs but messages embedded in socio-clinical systems. Their meaning depends on whether they are open to feedback, contestation, and reinterpretation. Without such feedback, *double noise* solidifies into clinical misjudgment.

The generalization this case supports is therefore structural rather than epidemiological: it demonstrates how default pipelines can normalize erasure as a design pattern, not how frequently such erasure occurs in deployed clinical systems. The specific figures reported here – accuracy rates, false negative counts, and cluster profiles – are artifacts of a synthetic dataset and should not be read as estimates of real-world performance. What the case makes visible is a logic, not a rate: the structural conditions under which clinically decisive cases disappear from model outputs, and why feedback is the mechanism that could make them reappear.

Conclusion: Feedback ethics and shared judgment

Across the cases and analyses developed in this article, double noise emerges not as a rare malfunction but as a structural condition of AI-mediated judgment. When algorithmic distortions intersect with the variability of human judgment – and when no channel exists to contest or contextualize outputs – small errors accumulate into entrenched misreadings. Put differently: predictive failure is not only statistical; it is communicative.

Two lessons follow from the juxtaposition of Leo's narrative and the stroke case. First, the absence of feedback transforms uncertainty into finality: once a system's inference is taken as decisive, lived signals cannot re-enter the chain of interpretation. Second, conventional performance metrics incentivize the neglect of rare but clinically decisive cases, masking harm behind reassuring aggregates.

Leo's case crystallizes these dynamics. What was in reality a localized cancer – serious but not impairing cognition – was reframed by algorithmic dashboards as a narrative of psychological decline. This was not only a medical misclassification but also a cultural injustice: her choice to keep working, shaped by collective norms of discretion, became reinterpreted as unreliability. The absence of feedback foreclosed any possibility of reintroducing her lived signal into the process, turning temporary fluctuations into definitive exclusion. This underscores why feedback ethics is not simply about correcting technical errors, but about keeping interpretive channels open across medical, organizational, and cultural registers.

The stroke case complements this by showing how blind spots emerge even in pedagogical settings. The dataset itself was designed for educational purposes, specifically to illustrate the risks of false positives in prediction exercises. Yet the analysis reveals a paradox: even in a context explicitly oriented toward teaching critical awareness, structural distortions persist when feedback is absent. This underscores that failures are not confined to flawed datasets or reckless deployment, but are built into the very routines of preprocessing, modeling, and evaluation when dialogue is displaced by default pipelines.

Feedback ethics names the shift required to meet these risks. Rather than treating feedback as a late-stage fix, it reframes prediction as an ongoing exchange in which claims must be open to challenge, revision, and situated justification. Shared judgment, in turn, treats clinicians, patients, and systems as co-participants in meaning-making, not as isolated producers and consumers of scores.

Operationalizing this orientation entails concrete design and governance commitments:

- **Harm-sensitive evaluation**: Prioritize recall for critical classes, report subgroup performance, and calibrate thresholds to clinical costs rather than aggregate accuracy. In stroke prediction, for instance, this means optimizing for sensitivity over specificity and reporting false negative rates disaggregated by age, comorbidity, and glucose profile.
- **Contestation by design**: Provide structured avenues to annotate, appeal, and escalate decisions; require human review for high-stakes or low-confidence outputs. Concretely, this could mean flagging cases where model confidence falls below a clinical threshold and routing them to a second reviewer before any administrative or therapeutic decision is taken.
- **Bidirectional logging**: Record model rationales *and* counter-evidence from clinicians and patients; route these signals into auditable update and retraining loops. In practice, this requires electronic health record systems that allow clinicians to annotate disagreements with algorithmic outputs and link those annotations to model retraining cycles.
- **Uncertainty literacies**: Present calibrated probabilities, selective abstention, and coverage indicators; train users to interpret limits rather than infer certainty. This includes embedding uncertainty communication into clinical training programs, so that practitioners can distinguish between a model's high-confidence prediction and its actual reliability on edge cases.
- **Proxy-risk audits**: Monitor and constrain inference from behavioral proxies that reconstitute sensitive attributes; align consent, purpose limitation, and model use. In workplace and clinical settings alike, this means periodically auditing whether productivity or behavioral metrics are being used as de facto indicators of cognitive or affective states without explicit consent or clinical validation.

These commitments extend beyond clinical AI to adjacent domains – education, finance, and law – where prediction without feedback hardens uncertainty into governance.

This account has limits that deserve explicit acknowledgment. Leo's story is fictional and the stroke dataset is pedagogical; neither substitutes for empirical field deployment, and the conclusions they support are structural rather than epidemiological. The framework of feedback ethics and shared judgment remains conceptual: its practical efficacy in real clinical workflows has not been tested, and it is not yet known whether institutionalizing contestation mechanisms would alter error distributions, reduce false negatives,

https://doi.org/10.1017/chr.2026.10034 Published online by Cambridge University Press

or change organizational practice in measurable ways. The stroke analysis, while transparent about its synthetic nature, cannot support claims about the frequency or severity of erasure in deployed clinical systems. Furthermore, the two cases examined here – workplace surveillance and stroke prediction – represent only a narrow slice of the contexts in which double noise can operate.

Future work should address these gaps on several fronts. Empirical studies should evaluate feedback mechanisms in real clinical workflows, measuring their effects on error distributions and studying how contestation alters decision-making over time. The framework would also benefit from interdisciplinary application to contexts where algorithmic and epistemic noise converge with particular intensity – such as the triage of self-harm cases in emergency departments, where the asymmetry between measurable somatic injury and invisible psychosocial crisis creates precisely the conditions under which double noise and foreclosed emplotment produce the highest clinical and ethical stakes. More broadly, the framework proposed here is intended as a transferable analytical lens for studying emergent narrative processes in human–AI team interactions and computational systems: wherever heterogeneous signals are configured into actionable stories under conditions of epistemic constraint, the concepts of emplotment, unreliable narration, and feedback ethics offer tools for diagnosing where meaning is made, where it is lost, and what would be required to keep it revisable.

From signals to stories, the core claim is simple: reliable clinical AI depends less on optimizing scores than on sustaining dialogue. Feedback ethics names the refusal to let emplotment freeze into verdict (Ricoeur 1984) – it keeps uncertainty accountable, technically, clinically, and culturally, so that what is predicted can still be questioned and what is lived can still be heard. Shared judgment, in turn, treats this accountability not as a constraint on performance but as its condition: whenever AI systems shape high-stakes human futures, sustaining dialogue is performance. In narratological terms, this is the condition of *situatedness* (Herman 2009): the possibility that every story, however fragmented by algorithmic compression, can still be told differently.

Data availability statement. The stroke dataset analyzed in this article is a publicly available pedagogical dataset commonly used in machine learning education. Code is available from the author upon reasonable request.

Acknowledgements. The author acknowledges the support of the Ministry of Science and Culture of Lower Saxony through funds from the program zukunft.niedersachsen of the Volkswagen Foundation for the "CAIMed – Lower Saxony Center for Artificial Intelligence and Causal Methods in Medicine" Project (Grant No. ZN4257). The author also acknowledges the use of AI-assisted tools (Claude, Anthropic) for language editing during the preparation of this manuscript.

Author contributions. The author was solely responsible for conceptualization, methodology, data curation, analysis, visualization, and writing of this article.

Competing interests. The author declares none.

Ethical standards. The research meets all relevant ethical guidelines, including adherence to legal requirements of the study country. No human participants or sensitive clinical data were involved.

References

Ajunwa, Ifeoma, Kate Crawford, and Jason Schultz. 2017. "Limitless Worker Surveillance." *California Law Review* 105, no. 3: 735–76. https://doi.org/10.15779/Z38BR8MF94.

Ananny, Mike, and Kate Crawford. 2018. "Seeing without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability." *New Media & Society* 20, no. 3: 973–89. https://doi.org/10.1177/1461444816676645.

Barnlund, Dean C. 1970. "A Transactional Model of Communication." In *Foundations of Communication Theory*, edited by **C. David Mortensen**, 83–92. New York: Harper & Row.

Barrett, Lisa Feldman. 2017. *How Emotions Are Made: The Secret Life of the Brain*. Boston, MA: Houghton Mifflin Harcourt.

Booth, Wayne C. 1961. *The Rhetoric of Fiction*. Chicago: University of Chicago Press.

Bürgi-Tian, Jennifer, and Stephan Boehm. 2021. "The Abrupt Dismissal: Organizational Practices and Employee Experiences of Immediate Termination." *Journal of Organizational Behavior* 42, no. 8: 1084–102. https://doi.org/10.1002/job.2543.

Char, Danton S., Nigam H. Shah, and David Magnus. 2018. "Implementing Machine Learning in Health Care—Addressing Ethical Challenges." *New England Journal of Medicine* 378, no. 11: 981–3. https://doi.org/10.1056/NEJMp1714229.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. "Smote: Synthetic Minority over-Sampling Technique." *Journal of Artificial Intelligence Research* 16: 321–57. https://doi.org/10.1613/jair.953.

Doshi-Velez, Finale, and Been Kim. 2017. "Towards a Rigorous Science of Interpretable Machine Learning." Preprint, arXiv:1702.08608. https://arxiv.org/abs/1702.08608.

Ekman, Paul. 1992. "An Argument for Basic Emotions." *Cognition & Emotion* 6, nos. 3–4: 169–200. https://doi.org/10.1080/02699939208411068.

Ekman, Paul, and Wallace V. Friesen. 1969. "The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding." *Semiotica* 1, no. 1: 49–98. https://doi.org/10.1515/semi.1969.1.1.49.

European Parliament and Council of the European Union. 2016. "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation)." Official Journal of the European Union, L 119. https://eur-lex.europa.eu/eli/reg/2016/679/oj.

European Parliament and Council of the European Union (EU). 2024. "Regulation (EU) 2024/1684 of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)." Official Journal of the European Union, L 168, 12 June 2024. https://eur-lex.europa.eu/eli/reg/2024/1684/oj.

Foucault, Michel. 1977. *Discipline and Punish: The Birth of the Prison*. Translated by Alan Sheridan. New York: Pantheon Books.

Fricker, Miranda. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

Gandini, Alessandro. 2019. "Labour Process Theory and the Gig Economy." *Human Relations* 72, no. 6: 1039–56. https://doi.org/10.1177/0018726718790002.

Genette, Gérard. 1980. *Narrative Discourse: An Essay in Method*. Translated by Jane E. Lewin. Ithaca, NY: Cornell University Press.

Ghassemi, Marzyeh, Luke Oakden-Rayner, and Andrew L. Beam. 2021. "The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care." *The Lancet Digital Health* 3, no. 11: e745–50. https://doi.org/10.1016/S2589-7500(21)00208-9.

Greene, Daniel, Anna Lauren Hoffmann, and Luke Stark. 2019. "Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical AI." In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, edited by Tung Bui, 2122–31. Honolulu: University of Hawaii at Manoa. https://hdl.handle.net/10125/59651.

He, Haibo, and Edwardo A. Garcia. 2009. "Learning from Imbalanced Data." *IEEE Transactions on Knowledge and Data Engineering* 21, no. 9: 1263–84. https://doi.org/10.1109/TKDE.2008.239.

Hendrycks, Dan, Mantas Mazeika, and Thomas Dietterich. 2019. "Deep Anomaly Detection with Outlier Exposure." In *7th International Conference on Learning Representations (ICLR 2019)*, New Orleans, LA, May 6–9, 2019. OpenReview.net. https://arxiv.org/abs/1812.04606.

Herman, David. 2009. *Basic Elements of Narrative*. Chichester: Wiley-Blackwell.

Hooker, Sara, Nithum Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. 2021. "Moving Beyond" Algorithmic Bias Is a Data Problem"." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FACCT)*, 29–40. https://doi.org/10.1145/3442188.3445899.

Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein. 2021. *Noise: A Flaw in Human Judgment*. New York: Little, Brown Spark.

Kelly, Christopher J., Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. 2019. "Key Challenges for Delivering Clinical Impact with Artificial Intelligence." *BMC Medicine* 17, no. 1: 195. https://doi.org/10.1186/s12916-019-1426-2.

Machado, André, Catarina Silva, and João Gama. 2023. "Machine Learning Applications for Detecting Anomalies and Ensuring Data Integrity in Clinical Trials." In *Proceedings of the International Conference on Health Informatics (HEALTHINF)*, 210–9. Springer.

McStay, Andrew. 2018. *Emotional AI: The Rise of Empathic Media*. London: SAGE Publications.

Meehl, Paul E. 1954. *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Minneapolis, MN: University of Minnesota Press.

Mittelstadt, Brent D., Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. "The Ethics of Algorithms: Mapping the Debate." *Big Data & Society* 3, no. 2: 1–21. https://doi.org/10.1177/2053951716679679.

Nelson, Lynn Hankinson. 1993. "Epistemological Communities." In *Feminist Epistemologies*, edited by **Linda Alcoff and Elizabeth Potter**, 121–59. New York: Routledge.

Nünning, Ansgar. 1999. "Unreliable, Compared to What? Towards a Cognitive Theory of Unreliable Narration: Prolegomena and Hypotheses." In *Grenz ü Berschreitungen: Narratologie Im Kontext / Transcending Boundaries: Narratology in Context*, edited by **Walter Grünzweig and Andreas Solbach**, 53–73. Tübingen: Narr.

Obermeyer, Ziad, and Ezekiel J. Emanuel. 2016. "Predicting the Future—Big Data, Machine Learning, and Clinical Medicine." *New England Journal of Medicine* 375, no. 13: 1216–9. https://doi.org/10.1056/NEJMp1606181.

Orwell, George. 1949. *Nineteen Eighty-Four*. London: Secker & Warburg.

Piper, Andrew, Richard Jean So, and David Bamman. 2021. "Narrative Theory for Computational Narrative Understanding." In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, edited by **Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih,** 298–311. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Rajkomar, Alvin, Jeff Dean, and Isaac Kohane. 2019. "Machine Learning in Medicine." *New England Journal of Medicine* 380, no. 14: 1347–58. https://doi.org/10.1056/NEJMra1814259.

Ricoeur, Paul. 1984. *Time and Narrative*, Vol. 1. Translated by Kathleen McLaughlin and David Pellauer. Chicago: University of Chicago Press.

Saito, Takaya, and Marc Rehmsmeier. 2015. "The Precision-Recall Plot Is more Informative than the Roc Plot when Evaluating Binary Classifiers on Imbalanced Datasets." *PLoS One* 10, no. 3: e0118432. https://doi.org/10.1371/journal.pone.0118432.

Schramm, Wilbur. 1954. "How Communication Works." In *The Process and Effects of Mass Communication*, edited by **Wilbur Schramm**, 3–26. Urbana: University of Illinois Press.

Sennett, Richard. 1998. *The Corrosion of Character: The Personal Consequences of Work in the New Capitalism*. New York: W. W. Norton.

Shannon, Claude E., and Warren Weaver. 1949. *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press.

Shortliffe, Edward H., and Martin J. Sepúlveda. 2018. "Clinical Decision Support in the Era of Artificial Intelligence." *JAMA* 320, no. 21: 2199–200. https://doi.org/10.1001/jama.2018.17163.

Topol, Eric J. 2019. "High-Performance Medicine: The Convergence of Human and Artificial Intelligence." *Nature Medicine* 25, no. 1: 44–56. https://doi.org/10.1038/s41591-018-0300-7.

Weir, Peter. 1998. *The Truman Show*. Film, Paramount Pictures.

Wiens, Jenna, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X. Liu, Finale Doshi-Velez, Kenneth Jung, et al. 2019. "Do no Harm: A Roadmap for Responsible Machine Learning for Health Care." *Nature Medicine* 25, no. 9: 1337–40. https://doi.org/10.1038/s41591-019-0548-6.

Wynants, Laure, Ben Van Calster, Gary S. Collins, Richard D. Riley, Georg Heinze, Ewoud Schuit, Marc M. J. Bonten, et al. 2020. "Prediction Models for Diagnosis and Prognosis of Covid-19 Infection: Systematic Review and Critical Appraisal." *BMJ* 369: m1328. https://doi.org/10.1136/bmj.m1328.

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. London: Profile Books.

https://doi.org/10.1017/chr.2026.10034 Published online by Cambridge University Press