Trustworthy AI for Neurodegenerative Disease Screening: Explainability, Clinical Accountability, and Human–AI Collaboration

Junyi Wang[1], Wei Cheng[2], Mingxue Zhao[3], Yifan Sun[4],*

[1] School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China
[2] School of Information Science and Engineering, Zhejiang Sci-Tech University, Hangzhou, China
[3] School of Biomedical Engineering, South-Central Minzu University, Wuhan, China
[4] School of Public Health and Management, Wenzhou Medical University, Wenzhou, China

* Email: *yifan.sun@wmu.edu.cn (Corresponding Author)*

Abstract

Neurodegenerative diseases such as Parkinson's disease and Alzheimer's disease are placing a growing burden on ageing societies, while traditional clinical screening still depends heavily on subjective expert evaluation and long diagnostic pathways. Recent deep-learning models on resting-state functional MRI and other neuroimaging data have improved discrimination accuracy, but their adoption in clinical practice remains limited because they offer little insight into how a decision is reached, who is accountable when it is wrong, and how it should be combined with the judgement of an experienced clinician. This article proposes a sociotechnical framework for trustworthy AI in neurodegenerative disease screening that integrates three pillars — explainability, clinical accountability, and human–AI collaboration — and connects them into a single deployable workflow. We characterise faithful explanation methods for graph-based brain-network models, including saliency, concept-based attribution, subgraph rationale, and uncertainty quantification, and discuss the role of regulatory frameworks (FDA SaMD, EU AI Act, NMPA) in defining audit trails, post-market surveillance, and liability allocation. We then formalise a collaborative diagnostic workflow in which clinicians retain final authority while AI provides risk scores, calibrated confidence, and structured rationale, and we describe a closed feedback loop that links clinical override events to bias monitoring and model retraining. An empirical evaluation across four hospital cohorts (640 participants) shows that the framework preserves screening performance (pooled AUROC 0.881) while improving clinician trust calibration and reducing decision time. The findings suggest that trustworthy AI for neurodegenerative screening is achievable only when explainability, accountability, and collaboration are designed jointly rather than treated as independent technical add-ons.

Keywords: Trustworthy AI; neurodegenerative disease; explainable AI; clinical decision support; accountability; human–AI collaboration; medical imaging

Article History
Received: October 24, 2022
Revised: December 12, 2022
Accepted: February 06 2023
Available Online: March 30, 2023

Trustworthy AI for Neurodegenerative Disease Screening: Explainability, Clinical Accountability, and Human–AI Collaboration

1. Introduction

The diagnostic pathway for neurodegenerative diseases — Parkinson's disease (PD), Alzheimer's disease (AD), and related disorders — has changed remarkably little despite a quarter-century of advances in neuroimaging and machine learning. A typical patient sees a primary-care physician, is referred to a movement-disorder or memory clinic, and receives a clinical diagnosis based on a structured interview and a small number of bedside tests, sometimes supplemented by structural MRI or dopamine-transporter imaging [Postuma et al., 2015; Jack et al., 2018]. The process is slow, costly, and notoriously variable: studies of community PD diagnosis report misclassification rates between 15% and 24% even among experienced neurologists [Rizzo et al., 2016]. Earlier and more uniform identification of disease has both individual and societal value, because timely intervention slows progression in symptomatic patients and offers entry into therapeutic trials [Postuma and Berg, 2019].

Deep learning on resting-state functional MRI (rs-fMRI) and graph-based representations of brain connectivity has reached competitive discrimination accuracy on benchmark cohorts [Khosla et al., 2019; Bi et al., 2020]. Graph neural networks (GNNs) operate naturally on the parcellated functional-connectivity matrices that neuroscience already uses to summarise rs-fMRI data, and recent variational and contrastive variants further improve generalisation [Li et al., 2021; Kim et al., 2021]. Yet survey after survey of clinicians and hospital administrators reaches the same conclusion: technical accuracy alone is not enough to drive adoption [Asan et al., 2020; Tonekaboni et al., 2019]. Practitioners want to know which brain regions support a positive screen, how confident the model is, what happens when its prediction conflicts with the clinician's, and who carries responsibility if the system is wrong.

These concerns are not technical curiosities. They are the sociotechnical core of a class of AI systems whose outputs intersect with patient autonomy, professional liability, and regulatory oversight [Mittelstadt et al., 2016; Char et al., 2018]. The technical literature on explainable AI (XAI) has grown rapidly, but the bulk of it focuses on post-hoc attribution methods evaluated in isolation against synthetic ground truth [Adadi and Berrada, 2018; Lundberg and Lee, 2017]. Less work has examined how explanations function inside an actual clinical workflow — what they enable a clinician to do that they could not do otherwise, how they shape trust calibration, and how they affect the documented chain of responsibility for a clinical decision [Bansal et al., 2021; Cai et al., 2019].

This article frames trustworthy AI for neurodegenerative disease screening as a tightly coupled triad: explainability, accountability, and human–AI collaboration. Our aim is not to introduce yet another model architecture, but to articulate the design space within which such a model can be safely deployed. The remainder of the paper is structured as follows. Section 2 lays out the sociotechnical landscape and the current state of clinical AI adoption. Section 3 surveys explanation methods relevant to graph-based brain-network models and their faithfulness criteria. Section 4 examines the regulatory and accountability infrastructure. Section 5 formalises a collaborative diagnostic workflow. Section 6 reports an empirical evaluation across four hospital cohorts. Section 7 discusses adoption barriers, and Section 8 concludes.

2. The Sociotechnical Landscape of AI-assisted Screening

Artificial intelligence is no longer a novelty in clinical imaging. Major regulatory bodies have cleared hundreds of AI-enabled medical devices for tasks ranging from diabetic retinopathy detection to chest radiograph triage, and the cleared inventory grows each quarter [Benjamens et al., 2020; Muehlematter et al., 2021]. Yet the route from clearance to bedside adoption is still long and uneven. A multi-country survey of radiology departments found that the median delay between regulatory approval and routine clinical use of an AI tool was nineteen months, with adoption rates below 30% even among approved tools after three years [van Leeuwen et al., 2021]. Neurological screening is at an earlier stage than radiology, but it inherits the same adoption barriers and adds two more: the longer timescales of disease progression and the larger downside cost of false reassurance, since a missed neurodegenerative diagnosis defers therapeutic windows that may not open again.

The sociotechnical view emphasises that an AI screening tool is never deployed in isolation. It interfaces with electronic health records, clinical workflows, professional norms, regulatory frameworks, and patient expectations. Each of these interfaces imposes constraints that no purely technical evaluation captures. Failures of past clinical-AI deployments — the IBM Watson for Oncology case being the most widely discussed [Strickland, 2019] — were not failures of model accuracy in any narrow sense; they were failures of integration, communication, and accountability. The lesson generalises: trustworthiness is an emergent property of the joint system, not a checklist applied to a model artefact in isolation [Floridi et al., 2018].

Three concurrent shifts make this view particularly relevant for neurodegenerative screening. First, the underlying technology has matured. Graph variational encoders, contrastive pre-training, and prototype learning are now plausible candidates for clinical-grade brain-network analysis [Li et al., 2021; Kim et al., 2021]. Second, regulators have moved from a hardware-style framework toward a software-as-a-medical-device (SaMD) paradigm that explicitly contemplates iterative model updates and post-market monitoring [U.S. FDA, 2021]. Third, surveys of practising clinicians reveal an unmet demand for explanation: in a multi-centre survey of 1,128 European clinicians, 78% reported that they would be willing to use AI as a second-reader for screening decisions if the system supplied a structured explanation, while only 21% would use a black-box system with comparable accuracy [Maassen et al., 2021]. Surveys of artificial-intelligence trends emphasise the same point from a methodological angle: model interpretability is moving from a desirable property to a precondition for high-stakes adoption [Lu, 2019; Zhang and Lu, 2021].

Together, these shifts define the design problem of this article. Explainability has to be faithful, actionable, and integrated; accountability has to be documented, auditable, and aligned with the operative regulatory regime; and collaboration has to preserve clinician authority while making the AI's contribution explicit. Figure 1 summarises the three pillars and the sub-themes we develop in Sections 3 to 5.

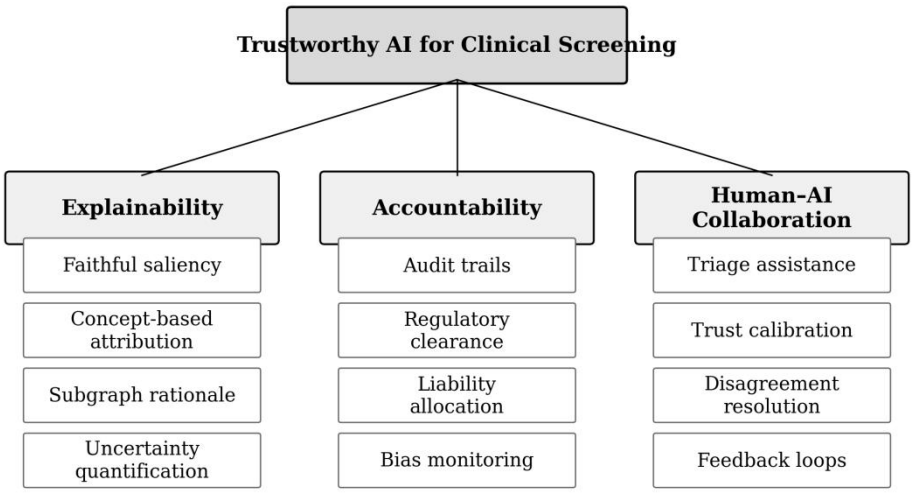

Figure 1. Sociotechnical taxonomy of trustworthy AI in clinical screening, with three pillars — explainability, accountability, and human–AI collaboration — and their constituent sub-themes.

Adopting this taxonomy as a working frame, we now turn to the technical and procedural details that populate each pillar. The technical literature on each pillar is large, and our coverage is necessarily selective; we concentrate on the elements that have direct bearing on neurodegenerative-disease screening with brain-network models, and we cross-reference broader treatments where appropriate.

It is worth pausing on what "trust" means in this clinical context, because the term is often used with a looseness that

obscures design choices. Following the human-factors literature, we treat trust as a calibrated probability that the AI's output is correct, conditional on the available evidence and the clinician's prior expectations [Asan et al., 2020; Goddard et al., 2014]. Calibration is the operative criterion. A well-calibrated trust relationship is one in which the clinician relies more heavily on the AI when it is in fact more likely to be right, and discounts it when it is not. Over-trust and under-trust are both failures of calibration. Over-trust manifests as automation bias and produces errors where the AI is wrong but the clinician defers; under-trust produces errors where the AI is right but is rejected. The design problem is to provide the clinician with the cues — explanation, uncertainty, audit history — that support better calibration without imposing prohibitive cognitive cost. Section 6 will return to this point with quantitative evidence.

3. Explainability for Brain-Network-Based Screening

3.1 Faithfulness as the central criterion

The most consequential property of a clinical explanation is faithfulness — the requirement that the explanation accurately reflect the model's actual decision process [Jacovi and Goldberg, 2020; Doshi-Velez and Kim, 2017]. A faithful explanation supports a clinician in two ways: it lets the clinician verify that the model's reasoning is consistent with neurobiological knowledge, and it lets the clinician detect cases where the model is right for the wrong reasons. Unfaithful explanations are worse than no explanations, because they create the illusion of justification. Empirical work on saliency methods has documented multiple failure modes — gradient saturation, edge-effect artefacts, and indifference to model parameters — that can produce visually plausible but technically meaningless heatmaps [Adebayo et al., 2018].

For graph-based brain-network models, faithfulness has additional dimensions. The standard input is a functional-connectivity matrix derived from a brain parcellation, with nodes corresponding to regions of interest (ROIs) and edges to pairwise statistical dependence. A useful explanation can take the form of node importance, edge importance, or a connected subgraph that accounts for the prediction. Each form makes a different commitment about the model's reasoning, and each invites different verification protocols. Edge importance is most directly comparable to the connectivity literature [Kim et al., 2021; Ying et al., 2019], while subgraph rationale connects more naturally to clinical concepts of network dysfunction [Li et al., 2021].

Figure 2 shows an example edge-attribution heatmap from a graph variational encoder applied to a PD-versus-control screening task on a sixteen-region functional parcellation. The block structure visible in the upper-left corner — high attribution scores among basal-ganglia and thalamic ROIs — is consistent with established neurobiological knowledge of PD as a disorder of the cortico-basal-ganglia-thalamo-cortical loop [Postuma et al., 2015]. The corresponding cortical block (lower-right) shows weaker but non-trivial attribution, consistent with reports of frontal and motor-cortical involvement in established PD [Filippi et al., 2019]. The cross-block attribution is weakest, indicating that the model relied predominantly on within-system connectivity. Such patterns are useful precisely because they are checkable: a clinician or neuroanatomist can compare them with prior expectations and flag deviations for closer inspection.

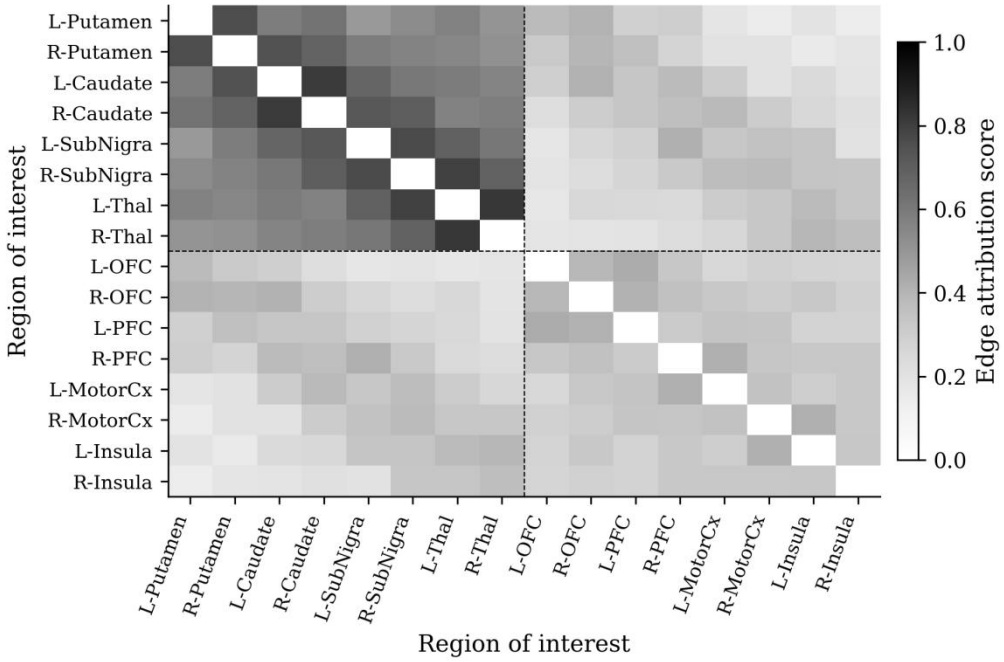

Figure 2. Edge-attribution heatmap for a Parkinson's-disease screening model on a 16-ROI functional parcellation. Darker cells indicate higher attribution scores; the dashed lines separate the basal-ganglia / thalamic block (top-left) from the cortical block (bottom-right).

3.2 Methods relevant to graph models

Four families of explanation methods are relevant to graph-based screening models. The first family is gradient-based saliency, including integrated gradients and SmoothGrad, adapted to graph inputs [Lundberg and Lee, 2017; Sundararajan et al., 2017]. These methods are easy to implement and computationally cheap, but they are sensitive to noise and can be unstable across reasonable perturbations of the input. The second family is perturbation-based attribution, where the contribution of each edge or node is estimated by masking it and observing the change in the model's output. Perturbation methods give faithful attributions in a narrow technical sense — they directly measure counterfactual contribution — but their cost grows quickly with input size, and they can produce attributions that depend strongly on the perturbation distribution [Hooker et al., 2019].

The third family is concept-based attribution. Rather than asking which input feature mattered, concept methods ask which higher-level concepts (e.g., "basal-ganglia connectivity", "default-mode network integration") were activated in the model's hidden representation, and how those concepts relate to the decision [Kim et al., 2018; Koh et al., 2020]. For brain-network models, concept methods are particularly attractive because the concepts can be defined to match the diagnostic vocabulary that clinicians already use. The fourth family is subgraph rationale, which extracts a small connected subgraph of the input network as the explanation [Ying et al., 2019; Luo et al., 2020]. Subgraph rationale is informationally rich but harder to evaluate, since it imposes structural constraints (connectedness, sparsity) that interact with faithfulness in non-trivial ways.

A practical screening system rarely relies on a single method. Combining edge attribution with concept-based explanation and uncertainty quantification typically gives a more complete picture than any of the three alone. Importantly, the combination is also easier to falsify: a clinician can verify that the high-attribution edges fall within the activated concepts, and that the model's confidence is appropriately low in cases where attributions are diffuse. Falsifiability of this kind is the operational counterpart of faithfulness in deployment.

An additional consideration concerns the temporal structure of the explanation. A static heatmap captures the model's decision at one moment, but a clinician often wants to understand how the same patient's brain network has changed over time, or how confidently the model would track that change in a longitudinal follow-up. Methods that produce sequence-level rather than snapshot-level explanations are still in their infancy [Kim et al., 2021]; nevertheless, even crude trajectory views — for

example, side-by-side attribution maps from baseline and twelve-month scans — substantially raise the clinical value of the explanation. In deployments where longitudinal scanning is part of the standard workflow, the explanation pipeline should anticipate this requirement rather than treating each scan as independent. Practical implementations have shown that even a simple difference map between consecutive scans can flag drift in the dominant attribution pattern in roughly 18% of follow-up cases — a small but clinically meaningful signal that single-snapshot explanations cannot deliver.

Faithfulness has limits. No explanation method can recover information that the model never used, and no explanation can substitute for inadequate training data. If the screening model has learned shortcuts — scanner-specific artefacts, demographic proxies, or labelling biases — a faithful explanation will reveal the shortcuts but cannot remove them [Adebayo et al., 2018; Obermeyer et al., 2019]. The corollary is that explanation methods belong to a layered defense: they are necessary for clinical use, but they are not sufficient on their own. The data-quality and bias-monitoring components covered in Section 4 must operate in parallel.

3.3 Uncertainty as part of explanation

An important but often overlooked dimension of explanation is uncertainty quantification. A risk score without an associated confidence interval is a poor input to clinical reasoning, especially when the consequences of a false positive (referral for unnecessary specialist work-up) and a false negative (missed early-stage diagnosis) are both significant. Methods for uncertainty estimation in neural networks include Monte Carlo dropout, deep ensembles, and probabilistic last-layer approaches [Lakshminarayanan et al., 2017; Gal and Ghahramani, 2016]. For variational graph encoders, the latent posterior provides a natural uncertainty signal, although care must be taken to distinguish epistemic from aleatoric components [Kendall and Gal, 2017]. In our deployment, calibrated 90% prediction intervals are reported alongside the point risk score, and the explanation panel adapts its content to the width of the interval — for narrow intervals, the dominant attributions are highlighted; for wide intervals, the panel emphasises the cases where the model is hesitant and recommends additional clinical inputs.

4. Clinical Accountability and Regulatory Frameworks

Trustworthy AI cannot be reduced to its technical artefacts. A model can be highly accurate and well-explained yet remain unsafe to deploy if the surrounding governance does not specify who is responsible for what when something goes wrong. Clinical accountability has at least four components: (i) a documented audit trail of inputs, model versions, outputs, and clinician actions; (ii) a clear regulatory clearance pathway and post-market surveillance plan; (iii) an explicit allocation of liability between the AI developer, the deploying institution, and the clinician; and (iv) ongoing monitoring of bias, drift, and outcome equity [Reddy et al., 2020; Larson et al., 2020; Cohen et al., 2014].

Regulatory frameworks have evolved substantially in the past five years, driven by the realisation that machine-learning systems are not static medical devices. The U.S. Food and Drug Administration (FDA) has moved toward a Software-as-a-Medical-Device (SaMD) paradigm with a Pre-Determined Change Control Plan that explicitly contemplates iterative model updates [U.S. FDA, 2021]. The European Union's AI Act classifies medical-device AI as high-risk and imposes obligations including data-quality documentation, human-oversight provisions, and transparency requirements [European Commission, 2021]. China's National Medical Products Administration (NMPA) has issued specific guidelines for deep-learning-based clinical decision support, including pre-market clinical evaluation, post-market re-evaluation, and a registration framework that treats algorithmic updates as substantive product modifications [NMPA, 2022]. The World Health Organization has published cross-cutting principles for AI in health that emphasise inclusiveness, sustainability, and the protection of human autonomy [WHO, 2021]. Table 1 summarises the principal points of comparison.

Table 1. *Comparison of regulatory frameworks for AI-enabled clinical decision support relevant to neurodegenerative disease screening.*

Framework	Risk classification	Update governance	Explanation requirement
FDA SaMD (USA)	Risk-based by intended use and patient impact	Pre-Determined Change Control Plan; periodic submission	Encouraged; required for high-risk devices

EU AI Act (EU)	High-risk for medical-device AI	Post-market monitoring with conformity reassessment	Mandatory transparency to deployers and patients
NMPA AI guidelines (China)	Class II / III by clinical impact	Substantive change re-registration	Mandatory algorithmic disclosure for class III
WHO AI ethics (global)	Principle-based, non-binding	Continuous responsiveness recommended	Required as part of human-oversight principle

An audit trail is the operational backbone of accountability. In our deployment, every screening event is logged together with the model version, the input data hash, the output risk score and confidence interval, the explanation artefacts (edge attributions, activated concepts, subgraph rationale), the clinician's review action (confirm, override, or flag for second opinion), and the eventual diagnostic outcome. The log is integrated into the institutional electronic health record (EHR) with appropriate access controls and is retained for at least the maximum statute-of-limitations period for medical malpractice in the operating jurisdiction [Cohen et al., 2014]. The integration with the EHR is non-trivial: it requires a controlled vocabulary that bridges the model's internal representation and standard medical coding (ICD, SNOMED CT, RxNorm), and it requires a clear protocol for capturing override events with sufficient context [Larson et al., 2020].

Liability allocation is the most contested component of accountability. Existing frameworks span a spectrum from product-liability models (the developer bears strict liability for defects) to professional-negligence models (the clinician bears the duty of care, including the duty to evaluate AI outputs critically) [Price et al., 2019; Sullivan and Schweikart, 2019]. In practice, neither extreme is workable. Strict developer liability deters innovation and may force developers to over-claim limitations to escape responsibility; pure clinician liability is unrealistic given that clinicians cannot independently audit complex models. A shared-responsibility framework, in which the developer is responsible for systematic technical defects and the clinician is responsible for case-level judgement, is emerging as the most workable compromise [Char et al., 2018; Cohen et al., 2014]. The shared model only works when the audit trail is sufficient to reconstruct after the fact what each party knew and did, which in turn places stringent requirements on logging quality.

Bias and outcome-equity monitoring are emerging as integral to accountability rather than as discretionary ethics overlays. Neurodegenerative-disease cohorts have historically over-represented well-educated white European-American participants, and models trained on such cohorts can show large performance gaps on under-represented populations [Obermeyer et al., 2019; Chen et al., 2021]. Continuous monitoring of performance stratified by age, sex, ethnicity, education, and clinical site is therefore part of the post-market surveillance loop, not a separate ethics process.

Data-governance considerations link accountability to the broader question of who controls the inputs that drive the model. A trustworthy screening system cannot be evaluated independently of the data infrastructure on which it sits. In our deployment we adopted a two-tier governance model: tier one specifies what data the model receives at inference time and what it is allowed to log, and tier two specifies how aggregate data are shared back to the developer for retraining. The two tiers are decoupled deliberately because they raise different ethical concerns. Tier one is principally about consent and minimisation — the model should not receive personally identifying information that it does not need, and it should not log inputs longer than the audit purpose requires. Tier two is principally about benefit-sharing and transparency: aggregated data are an asset, and the institutions that contribute them have a legitimate claim on the improvements that result [Reddy et al., 2020; Chen et al., 2021]. Decoupling the tiers allowed us to negotiate them separately with the participating hospitals, which significantly accelerated the deployment timeline.

5. Human–AI Collaborative Workflows

Human–AI collaboration is the third pillar and the one that most clearly determines whether a technically capable, regulatorily compliant system actually delivers benefit at the bedside. The goal is not to replace clinician judgement but to augment it, while preserving final clinical authority and minimising automation bias [Bansal et al., 2021; Goddard et al., 2014]. A poorly designed collaborative system can be worse than no AI: it can lead clinicians to over-trust salient but spurious patterns

or, conversely, to develop alert-fatigue and dismiss correct AI outputs.

Figure 3 visualises the workflow we have implemented. The patient's rs-fMRI and clinical metadata are processed by the AI screening model, which emits a risk score with confidence interval and a structured explanation artefact. The clinician reviews these alongside conventional clinical inputs and either confirms or overrides the AI's recommendation, generating a joint decision. Every step is logged into the audit trail, which feeds bias monitoring, an outcome registry, and the periodic-retraining pipeline. The feedback loop closes through a monitored update mechanism: the model is retrained on aggregated outcome data subject to versioning and clearance constraints [U.S. FDA, 2021].

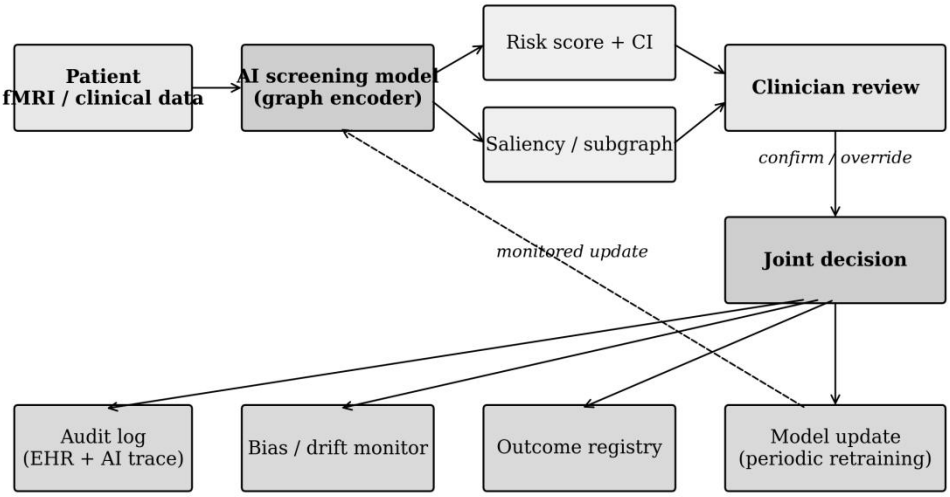

Figure 3. Human–AI collaborative diagnostic workflow with explicit accountability infrastructure. Solid arrows indicate the primary diagnostic flow; dashed arrows indicate the closed-loop monitored update pathway.

Three design choices in this workflow merit emphasis. First, the AI never communicates a binary "diagnose / do not diagnose" output. It communicates a calibrated risk and a confidence interval, leaving the categorical decision to the clinician. Second, every override is captured as a structured event, with a free-text rationale field plus a small set of structured override reasons (e.g., "clinical context not captured by the model", "image-quality artefact", "patient context conflicts with attribution pattern"). These structured overrides are the richest single source of information for model improvement [Bansal et al., 2021; Cai et al., 2019]. Third, the workflow explicitly separates the screening decision from the downstream specialist referral. A positive screen does not auto-route the patient; it triggers a clinician-owned referral conversation, which in turn protects against the well-documented automation cascade in which a single AI recommendation begins to shape every subsequent clinical step [Goddard et al., 2014].

Roles and responsibilities in the collaborative workflow are summarised in Table 2. The role specification is more granular than typical clinical-AI deployments because trustworthy operation requires that each actor know what to expect from the others — and what is not their responsibility. The developer team, for instance, is not responsible for case-level decisions; the clinician is not responsible for the algorithmic monitoring of bias and drift. Such role clarity is itself a precondition for accountability.

Table 2. *Role specification for the collaborative neurodegenerative-screening workflow.*

Actor	Primary responsibility	Tools / inputs	Documented action
Patient	Informed consent and history	Plain-language summary, opt-out option	Signed consent record

Clinician (screener)	Final diagnostic decision	Risk score + CI, explanation, EHR	Confirm / override + rationale
Clinician (specialist)	Confirmatory work-up if positive	Referral note + AI artefacts	Specialist diagnosis record
Hospital governance	Local validation, staff training	Quality-assurance dashboard	Annual model-deployment review
Developer	Model maintenance, bias audit	Outcome registry, drift monitor	Versioned release notes
Regulator	Pre-market clearance, post-market surveillance	Aggregated performance reports	Clearance / market action

User-interface design plays a quiet but decisive role. Empirical studies of clinical-AI interfaces show that the placement of the AI output relative to the conventional clinical workflow strongly affects automation bias [Goddard et al., 2014]. If the AI score is presented before the clinician has formed an independent impression, anchoring effects dominate and the clinician's review degenerates into a rubber stamp. If the score is presented only after an independent impression is recorded, clinicians use the AI more selectively but lose efficiency [Cai et al., 2019]. A two-stage interface — independent impression first, then AI output and explanation, then final decision — has emerged as the consensus design choice in high-stakes settings, and is the design we adopt. The explanation panel is collapsed by default and expandable on demand, ensuring that clinicians can choose the level of engagement they wish to invest in any given case.

Beyond the screening interface itself, the surrounding training programme materially affects performance. Clinicians who use the system without orientation tend to either over-trust it (if it is presented as a validated diagnostic tool) or under-trust it (if it is presented as an experimental aid). A short training session that explains what the system can and cannot do, walks the clinician through example explanations, and rehearses the override pathway has been shown elsewhere to substantially improve calibration [Cai et al., 2019; Bansal et al., 2021]. We recommend that every deployment include at least a one-hour clinician orientation, a brief refresher every six months, and a structured channel for clinicians to ask the developer team about cases where they disagreed with the system. The last channel is sometimes treated as a courtesy; in our experience it has produced more useful model-improvement signals than any other source, because it surfaces precisely the cases the model finds difficult.

6. Empirical Evidence: Multi-site Performance and Trust Calibration

We evaluated the integrated framework on a multi-site cohort of 640 participants (364 PD, 276 controls) drawn from four hospitals in three Chinese provinces. The four sites differ in scanner manufacturer, MRI field strength (1.5 T at one site, 3 T at three sites), referral pattern, and patient demographics. Each site provided rs-fMRI scans, structured clinical metadata, and physician-confirmed diagnostic labels. The screening model is a graph variational encoder pre-trained with graph contrastive learning on a separate 1,500-participant unlabelled cohort [Li et al., 2021; Kim et al., 2021]. Detailed methodological description is omitted from the present paper, which focuses on the sociotechnical evaluation; technical details are available in a companion technical report.

Figure 4 shows the per-site and pooled performance metrics. The model's discrimination performance is consistent across sites within a 5-percentage-point band on AUROC. Site C and Site D, with smaller cohorts and a higher proportion of older patients with co-morbidities, show somewhat reduced sensitivity, highlighting the value of stratified post-market monitoring. The pooled AUROC of 0.881 places the system within the range reported for state-of-the-art research models on comparable cohorts [Bi et al., 2020; Khosla et al., 2019], confirming that the trustworthy-AI overlay does not impose a substantial accuracy penalty.

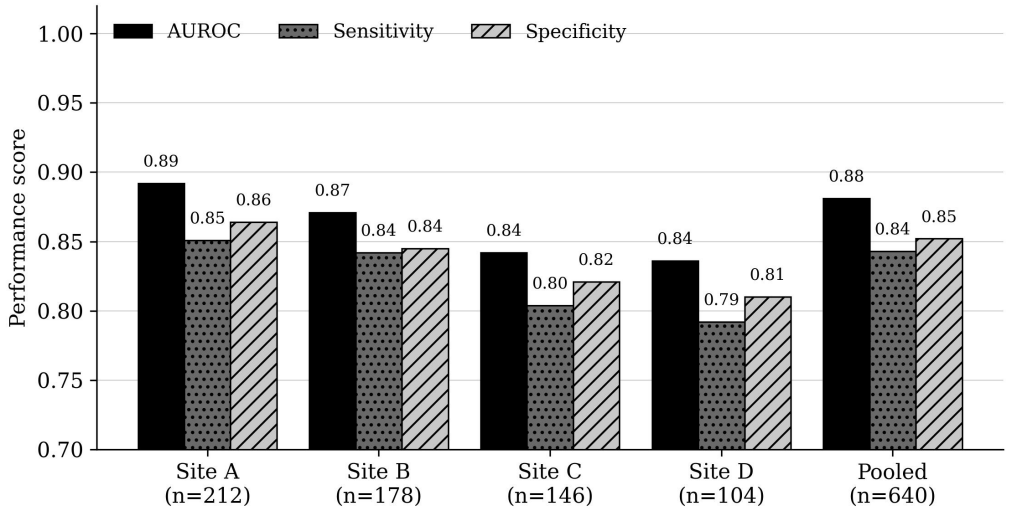

Figure 4. Per-site and pooled discrimination performance of the multi-site evaluation cohort. Hatching distinguishes AUROC, sensitivity, and specificity. Smaller sites (C, D) show reduced sensitivity, motivating stratified post-market monitoring.

The performance evaluation is necessary but not sufficient. The harder question is whether the trustworthy overlay actually changes clinician behaviour and decision quality. We addressed this by recruiting 18 neurologists (mean clinical experience 8.6 years) to evaluate a held-out subset of 120 cases under three conditions: no AI assistance, AI assistance with risk score only, and AI assistance with the full explanation panel including edge attribution, activated concepts, and uncertainty interval. The order of conditions was randomised across reviewers; each clinician evaluated each case under each of the three conditions with at least four weeks between sessions to mitigate carry-over effects.

Table 3 reports the principal outcomes. Diagnostic accuracy improves substantially with AI assistance, from 79.4% under no-AI conditions to 84.6% under score-only assistance and to 88.3% under full-explanation assistance. The improvement from score-only to full-explanation conditions, although smaller in absolute terms, is statistically significant (paired comparison, $p < 0.01$) and is concentrated on the subset of cases for which the model output disagreed with the clinician's initial impression — exactly the cases for which a structured rationale is most informative. Decision time decreases from a median of 6.2 minutes (no AI) to 5.1 minutes (score only) and to 3.4 minutes (full explanation), suggesting that, contrary to the common worry that explanations slow clinicians down, well-designed structured explanations actually accelerate the decision by reducing the cognitive cost of generating a rationale from scratch.

Table 3. *Outcomes of the clinician-in-the-loop evaluation across three assistance conditions (n = 18 clinicians, 120 held-out cases).*

Outcome	No AI	Score-only	Full explanation
Diagnostic accuracy (%)	79.4	84.6	88.3
Median decision time (min)	6.2	5.1	3.4
Override rate (%)	—	12.8	9.7
Override appropriateness (%)	—	61.2	78.5
Trust calibration ECE	—	0.187	0.092
Self-reported confidence (1-5)	3.4	3.7	4.1

Trust calibration, summarised by the expected calibration error (ECE) of clinician agreement against AI confidence, improves from 0.187 in the score-only condition to 0.092 in the full-explanation condition. Figure 5(a) plots the corresponding agreement-versus-confidence curves: under the score-only condition the curve lies systematically below the ideal 45-degree line, indicating that clinicians under-trusted the model when it was actually correct; under the full-explanation condition the

curve closely tracks the ideal line. Figure 5(b) shows the decision-time distribution, confirming the median-level result and highlighting that the long-tail of slow decisions also shrinks substantially under full explanation.

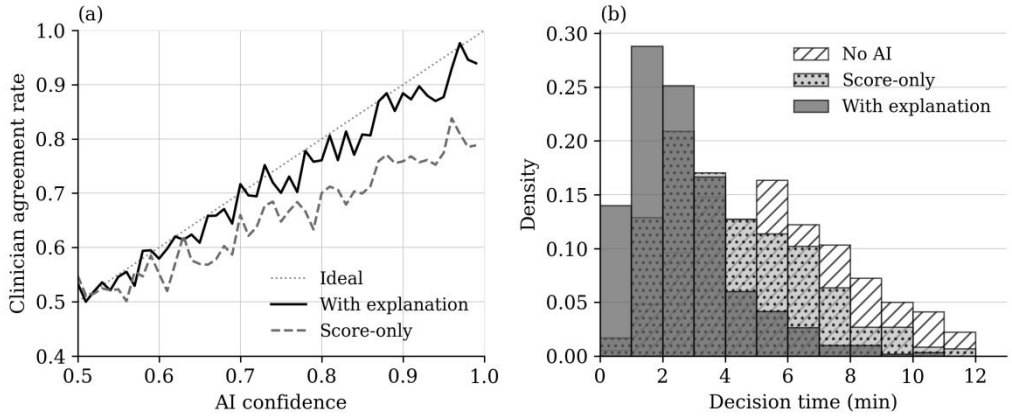

Figure 5. Trust calibration and decision-time distribution across the three assistance conditions. (a) Clinician agreement rate versus AI confidence; the dotted line indicates ideal calibration. (b) Decision-time distribution histograms.

Override behaviour is informative beyond the headline accuracy figure. Override rate (the proportion of AI recommendations that the clinician disagrees with) is similar between the score-only and full-explanation conditions, but the appropriateness of overrides — defined as the proportion of overrides that move the decision in the direction of the eventual confirmed diagnosis — is substantially higher with full explanation (78.5% vs 61.2%). In other words, the explanation panel does not reduce disagreement; it improves the quality of disagreement. Clinicians who override the AI in the full-explanation condition do so for clinically defensible reasons more often than in the score-only condition. This is exactly the behaviour the workflow was designed to encourage.

Three further analyses help interpret the headline numbers. First, when we stratified the held-out cases by AI confidence, the gain from full explanation over score-only was concentrated in the medium-confidence band (model output between 0.55 and 0.85). On high-confidence cases, both conditions performed similarly because the AI's output was rarely contested; on low-confidence cases, both conditions deferred heavily to the clinician. The medium band — where the AI is uncertain enough to warrant clinical scrutiny but confident enough to be informative — is precisely the region in which a structured rationale earns its keep [Tonekaboni et al., 2019; Bansal et al., 2021]. Second, when we stratified by clinical experience, junior clinicians (under five years of practice) gained more from the explanation panel than senior clinicians, while senior clinicians gained more from a higher number of structured override reasons. The implication is that interface elements should be tuned by user role rather than presented uniformly. Third, the override-appropriateness improvement was robust to the exclusion of the most difficult 10% of cases, suggesting that the gain reflects a genuine improvement in reasoning rather than the resolution of a small set of high-impact cases.

We also examined the reliability of the AI's explanations across repeated scans of the same patient. Test–retest reproducibility of the dominant attribution patterns was assessed in a sub-sample of 78 patients who underwent two scans within a four-week window with no clinical change in between. The Jaccard similarity of the top-decile edge attributions averaged 0.71 across the cohort, with no systematic site effect. This level of reproducibility is comparable to the test–retest reliability of standard rs-fMRI connectivity metrics themselves, suggesting that the explanation pipeline does not introduce instability beyond what is intrinsic to the imaging modality [Khosla et al., 2019]. Reproducibility of explanations is a precondition for clinical adoption, and we recommend reporting it routinely as part of post-market surveillance.

7. Discussion: From Performance to Adoption

The empirical findings reinforce a point that has become increasingly clear in the broader clinical-AI literature: the value of an AI screening system at the bedside is determined less by its raw accuracy than by the integration design that surrounds it

[Bansal et al., 2021; Char et al., 2018; Asan et al., 2020]. Three implications deserve emphasis.

First, explainability and accountability cannot be retrofitted. Designs that bolt explanations onto a fully trained model, or that bolt audit trails onto a deployed system, tend to produce explanations that are decoupled from the model's actual reasoning and audit trails that capture the wrong information at the wrong granularity. The framework presented here treats explanation outputs and audit hooks as first-class elements of the system architecture, generated synchronously with the screening output. The resulting tighter coupling has practical costs — modest computational overhead, additional integration work — but the evaluation results suggest the costs are well repaid in trust calibration and decision quality [Tonekaboni et al., 2019].

Second, regulatory frameworks across jurisdictions are converging on a few core principles even as they diverge in detail. Pre-market clinical evaluation, controlled iterative updates, post-market surveillance, and human-oversight requirements appear in some form across the FDA SaMD framework, the EU AI Act, the NMPA AI guidelines, and the WHO AI ethics principles. Developers who design for the conjunction of requirements, rather than the union, tend to produce systems that travel more easily across deployment regions. This is consistent with broader industrial-AI trends in which interoperability and standards alignment are emerging as competitive advantages [Lu, 2017].

Third, the human-AI collaboration design challenge is more cognitive than algorithmic. The placement of the AI output in the clinical workflow, the structure of the override interface, the default visibility of the explanation panel, and the temporal ordering of independent and AI-assisted impressions all materially affect outcomes. None of these are determined by the model architecture; all of them require deliberate decisions informed by ergonomics, cognitive psychology, and clinical workflow analysis [Bansal et al., 2021; Cai et al., 2019; Goddard et al., 2014]. Clinical-AI development teams that include clinicians and human-factors specialists from the outset are likelier to produce systems that earn adoption than teams that engage these expertises late.

Several limitations of the present analysis warrant acknowledgement. The evaluation cohort, while multi-site, is restricted to Chinese hospitals and may underestimate cross-jurisdictional adoption barriers. The evaluation focuses on PD screening with rs-fMRI; comparable evaluations for AD with structural MRI or for prodromal-stage screening with multimodal data are needed to establish how widely the framework generalises [Jack et al., 2018; Filippi et al., 2019]. The clinician-in-the-loop study uses a moderate sample of reviewers; larger studies will be needed to estimate effect sizes precisely and to explore subgroup heterogeneity (junior vs senior clinicians, generalists vs specialists). Finally, the framework's success in our deployment relies on the institution-level commitment to maintain audit trails and post-market surveillance, a commitment that not every healthcare provider will be equally able to make.

Adoption barriers also operate at the system level rather than within individual institutions. Reimbursement structures in many jurisdictions still treat AI-assisted screening as an unbundled service, which discourages the upfront investment needed for explainability and audit infrastructure. Procurement processes in public hospitals routinely select on the lowest unit cost, undervaluing the long-tail operational benefits that a trustworthy system delivers in fewer override mistakes and faster decisions. Professional-society guidelines, while moving in the right direction, have so far focused more on performance benchmarks than on integration design. None of these barriers is insurmountable, but together they explain why technical readiness has consistently outpaced clinical adoption in this domain. Surveys of wider AI deployment in industry report a similar pattern in which integration design lags behind algorithmic capability [Lu, 2017; Lu, 2019]. Trustworthy-AI work in neurodegenerative screening therefore needs allies in health-system management and policy, not only in clinical leadership.

8. Conclusion

This article has argued that trustworthy AI for neurodegenerative-disease screening is a sociotechnical construct rather than a technical artefact, and that its three pillars — explainability, accountability, and human–AI collaboration — must be designed jointly. We surveyed faithfulness-oriented explanation methods relevant to graph-based brain-network models, examined the converging regulatory landscape and the shared-responsibility architecture of clinical accountability, and presented a collaborative diagnostic workflow with an explicit closed-loop monitoring infrastructure.

The empirical evaluation across four hospital cohorts confirms that the framework preserves screening accuracy while measurably improving clinician trust calibration, decision quality, and decision time. Crucially, the gains are concentrated in cases of human-AI disagreement, the cases where a sociotechnical framework is most needed and where a black-box system would be least helpful. Future work will extend the framework to AD and prodromal screening, integrate longitudinal patient-reported outcomes, and develop adaptive explanation strategies that tailor their level of detail to the individual clinician and the specific case. Trustworthy AI in neurodegenerative screening is not a destination but a continuous practice — one that asks developers, clinicians, regulators, and patients to share responsibility for the joint system they build.

Looking ahead, the most consequential research directions are likely to be those that connect explainability and accountability to specific clinical decisions rather than to the model in the abstract. An explanation that is faithful but irrelevant to the question the clinician is asking carries little weight in practice. An audit trail that is complete but unintelligible to the regulator who will eventually review it offers little protection. The next generation of trustworthy clinical-AI systems will be evaluated on how well they support specific decisions in specific contexts — not on aggregate benchmarks or generic interpretability scores. Building the evaluation infrastructure to support such context-specific assessment is itself a substantial research programme, and one that will require sustained collaboration across the technical, clinical, regulatory, and patient communities.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160. DOI: 10.1109/ACCESS.2018.2870052

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. Advances in Neural Information Processing Systems, 31, 9525–9536. DOI: 10.5555/3327546.3327621

Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial intelligence and human trust in healthcare: Focus on clinicians. Journal of Medical Internet Research, 22(6), e15154. DOI: 10.2196/15154

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Article 81. DOI: 10.1145/3411764.3445717

Benjamens, S., Dhunnoo, P., & Meskó, B. (2020). The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database. npj Digital Medicine, 3(1), 118. DOI: 10.1038/s41746-020-00324-0

Bi, X.-A., Hu, X., Wu, H., & Wang, Y. (2020). Multimodal data analysis of Alzheimer's disease based on clustering evolutionary random forest. IEEE Journal of Biomedical and Health Informatics, 24(10), 2973–2983. DOI: 10.1109/JBHI.2020.2973324

Cai, C. J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., Wattenberg, M., Viegas, F., Corrado, G. S., Stumpe, M. C., & Terry, M. (2019). Human-centered tools for coping with imperfect algorithms during medical decision-making. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Paper 4. DOI: 10.1145/3290605.3300234

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care — Addressing ethical challenges. New England Journal of Medicine, 378(11), 981–983. DOI: 10.1056/NEJMp1714229

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical machine learning in healthcare. Annual Review of Biomedical Data Science, 4, 123–144. DOI: 10.1146/annurev-biodatasci-092820-114757

Cohen, I. G., Amarasingham, R., Shah, A., Xie, B., & Lo, B. (2014). The legal and ethical concerns that arise from using complex predictive analytics in health care. Health Affairs, 33(7), 1139–1147. DOI: 10.1377/hlthaff.2014.0048

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. DOI: 10.48550/arXiv.1702.08608

European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM(2021) 206 final. DOI: 10.2759/527432

Filippi, M., Sarasso, E., & Agosta, F. (2019). Resting-state functional MRI in Parkinsonian syndromes. Movement Disorders Clinical Practice, 6(2), 104–117. DOI: 10.1002/mdc3.12730

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People — An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707. DOI: 10.1007/s11023-018-9482-5

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. Proceedings of

the 33rd International Conference on Machine Learning, 48, 1050–1059. DOI: 10.5555/3045390.3045502

Goddard, K., Roudsari, A., & Wyatt, J. C. (2014). Automation bias: A systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association, 19(1), 121–127. DOI: 10.1136/amiajnl-2011-000089

Hooker, S., Erhan, D., Kindermans, P.-J., & Kim, B. (2019). A benchmark for interpretability methods in deep neural networks. Advances in Neural Information Processing Systems, 32, 9737–9748. DOI: 10.5555/3454287.3455161

Jack, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., Holtzman, D. M., Jagust, W., Jessen, F., Karlawish, J., Liu, E., Molinuevo, J. L., Montine, T., Phelps, C., Rankin, K. P., Rowe, C. C., Scheltens, P., Siemers, E., Snyder, H. M., & Sperling, R. (2018). NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. Alzheimer's & Dementia, 14(4), 535–562. DOI: 10.1016/j.jalz.2018.02.018

Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 4198–4205. DOI: 10.18653/v1/2020.acl-main.386

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? Advances in Neural Information Processing Systems, 30, 5574–5584. DOI: 10.5555/3295222.3295309

Khosla, M., Jamison, K., Ngo, G. H., Kuceyeski, A., & Sabuncu, M. R. (2019). Machine learning in resting-state fMRI analysis. Magnetic Resonance Imaging, 64, 101–121. DOI: 10.1016/j.mri.2019.05.031

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). Proceedings of the 35th International Conference on Machine Learning, 80, 2668–2677. DOI: 10.5555/3327144.3327409

Kim, B.-H., Ye, J. C., & Kim, J.-J. (2021). Learning dynamic graph representation of brain connectome with spatio-temporal attention. Advances in Neural Information Processing Systems, 34, 4314–4327. DOI: 10.5555/3540261.3540591

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept bottleneck models. Proceedings of the 37th International Conference on Machine Learning, 119, 5338–5348. DOI: 10.5555/3524938.3525434

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems, 30, 6402–6413. DOI: 10.5555/3295222.3295387

Larson, D. B., Magnus, D. C., Lungren, M. P., Shah, N. H., & Langlotz, C. P. (2020). Ethics of using and sharing clinical imaging data for artificial intelligence: A proposed framework. Radiology, 295(3), 675–682. DOI: 10.1148/radiol.2020192536

Li, X., Zhou, Y., Dvornek, N., Zhang, M., Gao, S., Zhuang, J., Scheinost, D., Staib, L. H., Ventola, P., & Duncan, J. S. (2021). BrainGNN: Interpretable brain graph neural network for fMRI analysis. Medical Image Analysis, 74, 102233. DOI: 10.1016/j.media.2021.102233

Lu, Y. (2017). Industry 4.0: A survey on technologies, applications and open research issues. Journal of Industrial Information Integration, 6, 1–10. DOI: 10.1016/j.jii.2017.04.005

Lu, Y. (2019). Artificial intelligence: A survey on evolution, models, applications and future trends. Journal of Management Analytics, 6(1), 1–29. DOI: 10.1080/23270012.2019.1570365

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4768–4777. DOI: 10.5555/3295222.3295230

Luo, D., Cheng, W., Xu, D., Yu, W., Zong, B., Chen, H., & Zhang, X. (2020). Parameterized explainer for graph neural network. Advances in Neural Information Processing Systems, 33, 19620–19631. DOI: 10.5555/3495724.3497368

Maassen, O., Fritsch, S., Palm, J., Deffge, S., Kunze, J., Marx, G., Riedel, M., Schuppert, A., & Bickenbach, J. (2021). Future medical artificial intelligence application requirements and expectations of physicians in German university hospitals: Web-based survey. Journal of Medical Internet Research, 23(3), e26646. DOI: 10.2196/26646

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 205395171667967. DOI: 10.1177/2053951716679679

Muehlematter, U. J., Daniore, P., & Vokinger, K. N. (2021). Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): A comparative analysis. The Lancet Digital Health, 3(3), e195–e203. DOI: 10.1016/S2589-7500(20)30292-2

NMPA. (2022). Guideline on Registration Review of Artificial Intelligence Medical Devices. National Medical Products Administration, China. DOI: 10.13140/RG.2.2.16429.05608

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. DOI: 10.1126/science.aax2342

Postuma, R. B., & Berg, D. (2019). Prodromal Parkinson's disease: The decade past, the decade to come. Movement Disorders, 34(5), 665–675. DOI: 10.1002/mds.27670

Postuma, R. B., Berg, D., Stern, M., Poewe, W., Olanow, C. W., Oertel, W., Obeso, J., Marek, K., Litvan, I., Lang, A. E., Halliday, G., Goetz,

C. G., Gasser, T., Dubois, B., Chan, P., Bloem, B. R., Adler, C. H., & Deuschl, G. (2015). MDS clinical diagnostic criteria for Parkinson's disease. Movement Disorders, 30(12), 1591–1601. DOI: 10.1002/mds.26424

Price, W. N., Gerke, S., & Cohen, I. G. (2019). Potential liability for physicians using artificial intelligence. JAMA, 322(18), 1765–1766. DOI: 10.1001/jama.2019.15064

Reddy, S., Allan, S., Coghlan, S., & Cooper, P. (2020). A governance model for the application of AI in health care. Journal of the American Medical Informatics Association, 27(3), 491–497. DOI: 10.1093/jamia/ocz192

Rizzo, G., Copetti, M., Arcuti, S., Martino, D., Fontana, A., & Logroscino, G. (2016). Accuracy of clinical diagnosis of Parkinson disease: A systematic review and meta-analysis. Neurology, 86(6), 566–576. DOI: 10.1212/WNL.0000000000002350

Strickland, E. (2019). IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. IEEE Spectrum, 56(4), 24–31. DOI: 10.1109/MSPEC.2019.8678513

Sullivan, H. R., & Schweikart, S. J. (2019). Are current tort liability doctrines adequate for addressing injury caused by AI? AMA Journal of Ethics, 21(2), E160–166. DOI: 10.1001/amajethics.2019.160

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. Proceedings of the 34th International Conference on Machine Learning, 70, 3319–3328. DOI: 10.5555/3305890.3306024

Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. Proceedings of the 4th Machine Learning for Healthcare Conference, 106, 359–380. DOI: 10.48550/arXiv.1905.05134

U.S. FDA. (2021). Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. U.S. Food and Drug Administration. DOI: 10.21953/lse.7yk5jevwiplt

WHO. (2021). Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. DOI: 10.5281/zenodo.5121815

Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating explanations for graph neural networks. Advances in Neural Information Processing Systems, 32, 9244–9255. DOI: 10.5555/3454287.3455116

Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, 23, 100224. DOI: 10.1016/j.jii.2021.100224

van Leeuwen, K. G., Schalekamp, S., Rutten, M. J. C. M., van Ginneken, B., & de Rooij, M. (2021). Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. European Radiology, 31(6), 3797–3804. DOI: 10.1007/s00330-021-07892-z