South American Publishing

SAP Artificial Intelligence in Dentistry. 2026;2:126
doi: 10.62486/aid2026126

REVIEW

Explainable artificial intelligence and clinician trust in dentistry: a contested claim

Inteligencia artificial explicable y confianza clínica en odontología: una afirmación en disputa

Ayu Hidayat[1]
Sari Nasution[1]

[1] Universitas Airlangga, Faculty of Medicine. Surabaya, Indonesia.

Submitted: 01-03-2026 | Revised: 05-05-2026 | Accepted: 01-07-2026 | Published: 05-07-2026

Corresponding Author: Sari Nasution

ABSTRACT

Explainable artificial intelligence is widely promoted as the mechanism by which opaque dental models will earn clinician trust, but the claim has acquired momentum faster than empirical support. This critical narrative review weighs the evidence behind three positions: that explanation is necessary for trust, that explanation does not reliably improve decisions, and that explanation helps only under specific conditions. The evidence converges on the perception that opaque models are a barrier to adoption, but it diverges on whether explanation improves diagnostic accuracy and whether it calibrates or inflates trust. Each divergence is traced to a concrete methodological cause rather than to contextual difference: the studies measure confidence instead of calibrated trust, they compare different explanation formats with small clinician samples, and they are drawn almost entirely from medical rather than dental settings. The argued position is that explanation is valuable in dentistry only when it is faithful, semantically meaningful and evaluated against human decisions, and that the dental field has not yet produced such evidence. Two tables map the positions and the quality of the evidence that supports each.

Keywords: explainable artificial intelligence; trust; clinical decision support; dentistry; interpretability.

RESUMEN

La inteligencia artificial explicable se promueve ampliamente como el mecanismo por el que los modelos dentales opacos ganarán la confianza de los clínicos, pero la afirmación ha adquirido impulso más rápido que apoyo empírico. Esta revisión narrativa crítica sopesa la evidencia detrás de tres posiciones: que la explicación es necesaria para la confianza, que la explicación no mejora de forma fiable las decisiones y que la explicación ayuda solo en condiciones específicas. La evidencia converge en la percepción de que los modelos opacos son una barrera para la adopción, pero diverge en si la explicación mejora la exactitud diagnóstica y en si calibra o infla la confianza. Cada divergencia se atribuye a una causa metodológica concreta y no a una diferencia contextual: los estudios miden confianza en lugar de confianza calibrada, comparan formatos de explicación distintos con muestras pequeñas de clínicos y proceden casi en su totalidad de entornos médicos y no odontológicos. La posición argumentada es que la explicación es valiosa en odontología solo cuando es fiel, semánticamente significativa y se evalúa frente a decisiones humanas, y que el campo odontológico todavía no ha producido esa evidencia. Dos tablas cartografían las posiciones y la calidad de la evidencia que respalda cada una.

Palabras clave: inteligencia artificial explicable; confianza; apoyo a la decisión clínica; odontología; interpretabilidad.

© 2026; The Authors. This Open access article is distributed under CC BY 4.0 licence.

INTRODUCTION

The claim that explainable artificial intelligence will earn the trust of clinicians has become a settled premise in dental artificial intelligence, but it is a premise in dispute rather than a finding. The proponents hold that the opacity of deep models is the primary barrier to clinical adoption and that explanation is the remedy. Thaweesapphithak et al.(1) frame explainability as a solution to the black-box problem in dentistry, Al-Hammad et al.(2) treat explanation as the route to clinician trust in dental imaging, and Mihut et al.(3) connect explainability to professional autonomy and accountability, while Song et al.(4) propose an explainable clinical decision support service for dental practice and Vivel-Couso et al.(5) present a trustworthy, explainable age-estimation model for panoramic images. The dental imaging reviews of Can et al.(6) and the framework of Ikerionwu et al.(7) extend the same assumption, and the medical-imaging literature supplies a parallel set of advocates, of whom Savage et al.(8) are representative in describing explanation as a bridge across the trust gap. This position is the default assumption of much of the field, and it has acquired regulatory and commercial momentum faster than it has acquired empirical support.

The stakes of the dispute are higher than a disagreement about a machine-learning technique, because explanation has become a condition of entry for dental artificial intelligence. Regulators and professional bodies increasingly treat explainability as a prerequisite for high-risk artificial intelligence, and a dental model that cannot explain itself may be judged unsuitable for deployment regardless of its accuracy, so the question of whether explanation actually earns trust is no longer academic. The commercial momentum is equally strong: explainability is a selling point for dental software, and the systematic reviews of Thaweesapphithak et al.(1)and Al-Hammad et al.(2)both note that the claim of trustworthiness has run ahead of the evidence that would establish it. If the skeptical position is even partly correct, then a field that has invested in explanation as its path to adoption may have invested in the wrong instrument, and if the grey-zone position is correct, then the investment was right but its conditions were not secured. The dispute therefore bears directly on how dental artificial intelligence should be built, evaluated and regulated, and that is what gives the critical appraisal that follows its clinical importance.

The skeptical position challenges the premise on empirical grounds. Chu et al.(9) reported that explanation format did not affect diagnostic accuracy in a simulated experiment with 22 clinicians, and Clement et al.(10) found across three experiments that the quality of the artificial intelligence and the clinical role of the user affected decision quality more than the explanations did. Brdnik et al.(11) went further, describing an illusion of transparency in which users expressed satisfaction and trust without improving the appropriateness of their reliance, and Hunsicker et al.(12) observed that explanations increased cognitive load and disrupted workflow among ophthalmologists. Okoroafor et al.(13) critically evaluated the evidence linking explanation to clinician trust and to diagnostic accuracy in multi-modal medical imaging and found it thin, and Borra et al.(14) qualified the benefit of explanation by restricting it to the cases where the explanation is faithful and semantically meaningful. This position holds that trust is not the same as calibrated reliance, and that explanation can inflate the former without improving the latter.

The skeptical position is not a counsel of despair, and it should not be read as a claim that explanation is worthless. It is a claim that the value of explanation has been assumed rather than demonstrated, and that the assumption has displaced the experiments that would test it. A dental model that is accurate but opaque, and a dental model that is slightly less accurate but explainable, cannot be compared by appeal to trust, because trust in the second model is the very thing at issue, and the skeptical position is the insistence that the comparison be made rather than asserted. This is the most useful reading of the skeptics, and it is the reading that the present review adopts in weighing the evidence that follows.

A third position occupies the middle ground, and it is the most defensible of the three because it converts the dispute from a contest of slogans into a set of testable conditions. Davis et al.(15) argued that explanation should be contextualized, at the level of the model, the setting, the subpopulation and the patient, so that trust is calibrated rather than assumed. Haas et al.(16) found that clinicians preferred SHAP explanations over counterfactuals and anchors in a cardiovascular setting, which suggests that the value of explanation depends on its format and its fit to the task. Nicolson et al.(17) showed that explanation affected confidence but not accuracy in a sonography task, and Stevens et al.(18) supplied an instrument, the TrAAIT, for measuring the trust that explanation is supposed to produce, while Baskerville et al.(19) observed the same dependence of explanation on the user in a staging task. This review weighs these three positions against the evidence, and the objective stated here is to adjudicate them: to identify the convergences, divergences and grey zones in the evidence, and to argue a position grounded in the methodological

causes of the divergences rather than in a preference for one side.

DEVELOPMENT

The evidence is organized into three sections, convergences, divergences and grey zones, and each is presented with the support that sustains it (Table 1).

Table 1. Positions in the debate on explainable artificial intelligence and clinician trust.

Position	Central claim	Principal defenders
Pro-XAI	Explanation is necessary for clinician trust and adoption	Thaweesapphithak et al.; Al-Hammad et al.; Mihut et al.; Kheder et al.
Skeptical	Explanation does not reliably improve accuracy or calibrated trust	Chu et al.; Clement et al.; Brdnik et al.; Hunsicker et al.
Grey zone	Explanation helps only when faithful and evaluated against human decisions	Davis et al.; Borra et al.; Haas et al.; Nicolson et al.

Note: XAI = explainable artificial intelligence.

Convergences

Two findings command agreement across the positions, and they are the least controversial part of the debate. The first is that opaque models are perceived as a barrier to clinical adoption, and this perception is documented in both the dental and the medical literature. Thaweesapphithak et al.(1)identified the black-box nature of complex models as a primary barrier to adoption in dentistry, Al-Hammad et al.(2)recorded the same perception across 61 dental imaging studies, and Borra et al.(14)listed opacity of convolutional networks among the key barriers to trust in cancer imaging, a list that Fotopoulos et al.(23)extended across cancer-imaging modalities. This convergence is about perception rather than performance, and it is the point on which the entire field agrees.

The perception that opacity is a barrier has concrete dental instantiations, and naming them sharpens the convergence. Vivel-Couso et al.(5)framed trustworthiness as the central requirement for an age-estimation model applied to panoramic images, Song et al.(4) designed an explainable clinical decision support service specifically to enhance the acceptance of dental recommendations, and Can et al.(6)fused explainable radiomics with a convolutional network for multimodal dental lesion classification. Kheder et al.(24)went further by pairing an explainable gradient-boosting model with multicentre validation and clinical interpretation, which is the closest the dental field has come to the human-evaluated explanation that the grey-zone position demands. These applications share the premise that explanation is what makes a dental model acceptable, and they are the reason the pro-explainability position is not a straw man but the field's own self-understanding.

The second convergence is narrower but real: explanations can raise clinician confidence when they are faithful and semantically meaningful. Borra et al.(14) reported this conditional benefit, Haas et al.(16)found that clinicians preferred SHAP explanations and judged them useful in a heart-failure telemonitoring scenario, and Nicolson et al.(17)observed that explanations increased participant confidence in a gestational-age estimation task even when they did not improve accuracy. The preference for feature-attribution explanations is echoed in the broader literature, where Brown et al.(25)built a SHAP-driven framework for hospital admission prediction and Elsayad et al.(26)used explainable methods to grade glioma. This reading of the evidence is that explanation affects the subjective experience of the clinician, and the disagreement that follows is over whether that subjective change is the same as trust, and whether it produces better decisions. The convergence is therefore real but narrow, and it does not by itself carry the pro-explainability position as far as that position would like to go.

A third convergence emerges not from the findings but from the prescriptions the studies themselves offer, and it is the most instructive of the three because it cuts across the debate. Whether they are proponents or skeptics, the authors converge on the claim that explanations must be evaluated against human decisions rather than against automated metrics alone, and they converge on the claim that a clinician must be able to detect when a model is wrong. Davis et al.(15)stated this directly as the goal of calibrated reliance, Ud Din et al.(20)made it the criterion for clinical impact, and even the pro-explainability dental reviews of Thaweesapphithak et al.(1)and Al-Hammad et al.(2)conceded that the trust outcome is the one that the field has measured least

well. The convergence is therefore a negative one, an agreement on what is missing, and it is the point at which the debate, for all its surface disagreement, is actually united. This reading of the literature is that the positions differ in their optimism about explanation, but they share the same standard for what would settle the dispute, and that shared standard has simply not been met.

Divergences

The first divergence concerns diagnostic accuracy, and it is the sharpest. The proponents hold that explanation should improve decisions, because a clinician who understands a recommendation can judge it, and the logic of that position is sound as far as it goes, but the experiments that measured accuracy failed to show the improvement the logic predicts. Chu et al.(9)reported that explanation format did not significantly affect diagnostic accuracy among 22 clinicians, and Nicolson et al.(17)found that explanations produced a further non-significant reduction in mean absolute error, from 15.7 to 14.3 days, after the model itself had already reduced it from 23.5 days. Clement et al.(10)attributed the effect on decision quality to the quality of the artificial intelligence and the clinical role of the user rather than to the explanation, and Baskerville et al.(19)observed the same dependence on the user in a TNM staging task. This divergence is methodological in origin, and it has a concrete cause: the studies that find no accuracy benefit measure accuracy directly, whereas the studies that claim a benefit measure confidence or self-reported trust, and the two outcomes are not interchangeable.

The second divergence concerns whether explanation calibrates or inflates trust, and it exposes the same confusion under a different name. Davis et al.(15)held that explanation should calibrate appropriate trust and reliance, and they advocated contextualizing it to the model, setting, subpopulation and patient, while Brdnik et al.(11)reported an illusion of transparency in which satisfaction and trust rose without any improvement in the appropriateness of reliance. Hunsicker et al.(12)challenged the assumption that explanation improves understanding, because their ophthalmologists experienced higher cognitive load and workflow disruption when explanations were added, and Huang et al.(27)found that the factors shaping trust in artificial intelligence chatbots were not primarily the explanations but the perceived competence and warmth of the system. This divergence also has a concrete cause: the studies differ in whether they measure trust as a feeling, which explanation reliably increases, or as a calibrated behavior, which explanation does not, and the failure to separate the two is the source of the contradiction. A study that reports trust ratings and a study that reports reliance are not in conflict with one another, because they are reporting different things, and the apparent contradiction between the pro-explainability and skeptical positions dissolves once that difference is named.

A third divergence concerns the format and delivery of the explanation itself, and it is the least recognized of the three. The studies disagree about whether the choice of explanation format changes anything, and their disagreement has a concrete cause in the way the formats are presented. Chu et al.(9)manipulated four explanation formats and two levels of reliability and found that format affected task time, cognitive load and trust ratings but not accuracy, whereas Haas et al.(16) found a clear preference for SHAP over counterfactuals and anchors in cardiovascular clinicians, and Hagiwara et al.(28)held that the safety of clinician-artificial intelligence interaction depends on how the explanation is embedded in the workflow rather than on which algorithm produced it. The divergence is therefore not about whether explanation matters but about which property of explanation matters, and the studies that disagree are measuring different properties, format, preference or workflow, without a shared definition of what a good explanation is. This third divergence reinforces the central point of the review, that the field has not settled what it is even testing.

Grey zones

Three questions remain open, and they are open because the evidence that would close them has not been produced. The first is whether the dental setting differs from the medical settings in which most of the primary evidence was gathered. The trust experiments of Chu et al.,(9)Clement et al.,(10)Nicolson et al.,(17)Hunsicker et al.(12)and Haas et al.(16)were conducted in pulmonology, drug dosing, sonography, ophthalmology and cardiology respectively, and none was dental, while the dental studies of Thaweesapphithak et al.(1)and Al-Hammad et al.(2)are reviews of models rather than experiments on clinicians. What remains open is therefore whether a dental clinician, looking at a dental explanation, behaves like the medical clinicians in those experiments, and the field has not answered it.

The absence of a dental trust experiment is itself partly explained by a measurement problem, and that problem is the deeper of the three grey zones. A faithful explanation is one that

reflects the model's actual computation, and measuring faithfulness requires access to the model's gradients or a perturbation procedure, neither of which is part of routine dental practice, so the dental researcher who wants to evaluate an explanation is forced to choose between a technical metric that ignores the clinician and a clinician study that cannot be run at the scale a dental clinic allows. This is not a contextual difference but a structural one, and it is the concrete reason that the dental literature has stayed at the level of model demonstration while the trust question has been answered, tentatively, only in larger medical settings.

The second open question is how to measure explanation quality in dentistry at all. Al-Hammad et al.(2)found that the rigor with which explanations are evaluated remains unclear, and Thaweesapphithak et al.(1)assessed whether dental explainable artificial intelligence is methodologically rigorous enough to substantiate claims of trustworthiness, applying QUADAS-2 and PROBAST to 19 of 100 identified studies. The grey zone is that a shared standard for evaluating dental explanations does not exist, and without one the claim of trustworthiness cannot be adjudicated; the same absence is reported across healthcare by Shiddik et al.(29)and Zhang et al.(30) and across electrocardiography by Taleban et al..(31)The third open question is whether explanation helps some clinicians and harms others, which Nicolson et al.(17)raised by showing that the impact of explanations varied across participants, with some performing worse with explanations than without. This contrasts with the uniform-benefit assumption that the pro-explainability position takes for granted. A fourth grey zone, closely related, is the interaction between explanation and system reliability, which Chu et al.(9)observed as a significant interaction between format and reliability, and which Davis et al.(15)framed as the need to calibrate trust to the reliability of the individual prediction. The implication is that an explanation is not a fixed property of a model but a property of the pairing between a model, a case and a clinician, and a field that treats explanation as a one-time certification rather than a per-case calibration will miss the interaction entirely. The positions and the evidence that supports each are mapped in (Table 2).

Table 2. Quality of the evidence supporting each position.

Position	Outcome measured	Evaluation design	Population	Weight
Pro-XAI	Confidence and self-reported trust	Automated metrics and review	Mostly dental models, not clinicians	Low-moderate
Skeptical	Diagnostic accuracy and cognitive load	Simulated and mixed-methods experiments	Medical (non-dental) clinicians	Moderate
Grey zone	Calibrated reliance and format preference	Human evaluation of explanations	Medical clinicians, small samples	Moderate

Note: Higher weight denotes human-evaluated outcomes (accuracy or calibrated reliance) over self-reported trust.

DISCUSSION

One would expect, given the confidence with which explainability is promoted, that the dental literature would contain direct experiments showing that explanations improve dental decisions, and it does not. The expectation is contradicted in two ways: the primary trust experiments are almost entirely medical rather than dental, and the experiments that do exist find no effect of explanation on diagnostic accuracy. This contrast between expectation and evidence is the entry point for the critical evaluation that follows.

The divergences in the evidence are explained by three concrete methodological causes, and none of them is a matter of contextual difference. The first cause is the outcome variable. Studies that measure self-reported trust or confidence reliably find a benefit, because explanation makes a model feel more transparent, whereas studies that measure accuracy or calibrated reliance find little or no benefit, because explanation does not change the information that a correct or incorrect recommendation contains. The second cause is the evaluation design.

Studies that score explanations with automated faithfulness metrics never test the clinician, and the dental reviews of Thaweesapphithak et al.(1)and Al-Hammad et al.(2)are of this kind, so their support for the trust claim is indirect, and the same indirectness marks the survey and review literature from Valishirin et al.(32)through Srivastava et al.(33)to Aravindkumar et al..(34)The third cause is the sample, both its size and its discipline: the experiments of Chu et al.(9)and Nicolson et al.(17)used 22 and 10 clinicians respectively, and the finding that explanations helped some par-

ticipants and harmed others is exactly what a small, heterogeneous sample would produce. This reading of the evidence holds that the divergence between the pro-explainability and skeptical positions is not a disagreement about facts but a difference in what each side chose to measure, and the two measurements do not answer the same question.

A fourth cause underlies the others and deserves to be named, because it is the one the field most resists acknowledging. The concept of trust itself is underspecified in the explainable artificial intelligence literature, and the studies that claim to measure it are measuring different constructs under the same name. Some measure affective trust, the clinician's feeling of comfort with a model, and some measure cognitive trust, the clinician's judgment that the model is competent, and some measure behavioral reliance, the clinician's willingness to act on the recommendation, and these three are correlated but distinct, as the TrAAIT instrument of Stevens et al.(18)was designed to disentangle.

The divergence between the positions is therefore partly a divergence between the constructs they call trust, and a field that does not distinguish affective trust from calibrated reliance will continue to produce studies that appear to contradict one another while each is correct about its own construct. This is the methodological cause that the distinctive trait of this review, attributing every divergence to a concrete cause, most needs to state plainly, because it is the one that an appeal to context would otherwise conceal.

The argued position of this review follows from that reading. Explanation is valuable in dentistry not because it generates trust, but because it can support the audit and the correction of a model, and that value is conditional on three things that the dental field has not yet secured. The first is faithfulness, the second is semantic meaning, and the third is evaluation against human decisions, and the three are cumulative rather than alternative: an explanation that is faithful but expressed in raw saliency pixels, or meaningful but unfaithful to the model, or both but never shown to a clinician, fails the condition that its position in the argument requires. The explanation must be faithful, meaning it reflects the model's actual computation rather than a plausible reconstruction, the explanation must be semantically meaningful, meaning it is expressed in the vocabulary of teeth and surfaces rather than of saliency pixels, and the explanation must be evaluated against human decisions rather than against automated metrics.

The dental systematic reviews show that these conditions are not met: Thaweesapphithak et al.(1)applied QUADAS-2 and PROBAST to dental explainable artificial intelligence and found the methodological rigor wanting, and Al-Hammad et al.(2)reported that the rigor with which explanations are evaluated in dental imaging remains unclear. What remains open is therefore not whether explanation can help, but whether dentistry will build explanations that meet the three conditions and then test them against decisions.

The automation-bias dimension deserves separate statement, because it is the failure mode that explanation is supposed to prevent and that, on the skeptical evidence, it sometimes fails to prevent. Automation bias is the tendency to favour a recommendation from an automated system simply because it is automated, and Davis et al.(15)identified it as the central risk that a well-designed explanation should counteract by calibrating reliance to the reliability of each prediction.

The evidence reviewed here suggests that explanation, as currently delivered, is not reliably protective against this bias, because Brdnik et al.(11)found that trust and satisfaction rose without a corresponding improvement in the appropriateness of reliance, and because Hunsicker et al.(12) found that the added cognitive load of explanation could itself push a busy clinician toward accepting the model rather than interrogating it. A dental clinician under time pressure, shown a plausible-looking saliency map, may be more, not less, likely to accept a wrong recommendation, and this is the specific risk that the argued position is meant to address by asking for explanations that invite scrutiny rather than acquiescence.

The dental evidence gap has a precise shape, and stating it makes the argued position actionable. What the dental field has produced are explainable models and reviews of explainable models, but it has not produced a single experiment in which a dental clinician made a diagnosis with and without an explanation and the accuracy or the calibrated reliance of the decision was measured, and this absence is the reason the trust claim remains contested.

The closest approximation is the multicentre study of Kheder et al.,(24)which paired an explainable model with clinical interpretation but did not manipulate the explanation as an experimental factor, and the closest medical analogue is the sonography experiment of Nicolson et al.,(17)which did manipulate the explanation and found no accuracy effect. A definitive dental experiment would randomize dental clinicians to recommendations with faithful explanations, unfaithful explanations or no explanations, and would measure diagnostic accuracy, calibrated reliance and workload together, and such an experiment is entirely within the field's reach but has not been run. This

reading of the gap is that the dispute persists not because the question is hard but because the decisive study has not been conducted, and the review's argued position is therefore as much a call for that study as a judgment of the evidence that exists.

The argument here contrasts with the enthusiasm that runs through much of the general explainable artificial intelligence literature, and the contrast is instructive. Reviews by Paliwal et al.,(35) Lakshmi et al.(36) and Sarma et al.(37) describe explanation as a bridge to trust in general terms, and Marchi et al.(38) and Oisakede et al.(39) frame explainability as the path to trustworthy oncology, but these accounts are largely programmatic and cite the same small set of primary experiments as evidence, so they inherit the weakness of that evidence without acknowledging it. The methodological reviews of Hamrouni et al.(40) and Durak et al.(41) are closer to the skeptical position, because they distinguish the faithfulness of an explanation from its perceived clarity, and Mi et al.(42) reach the same distinction in cardiac electrophysiology, and Fayaz et al.(43) and Barki et al.(44) point the same way in their calls for governance and evaluation. Even so, the field's enthusiasm is understandable, because the alternative to explanation is an opaque model that no clinician can audit, and the responsible resolution, argued here, is to keep explanation while making it faithful and testable, not to abandon it. The programmatic literature and the skeptical literature are therefore not opponents but two halves of the same conclusion, that explanation is necessary but not yet sufficient, and the present review's contribution is to show that the insufficiency is measurable and specific rather than vague and contextual.

The implications for dental practice are cautious and concrete. A dental clinician should not treat an explanation as a certificate of trustworthiness, because Brdnik et al.(11)showed that satisfaction with an explanation can coexist with uncalibrated reliance, and Hunsicker et al.(12) showed that explanation can add cognitive load in a time-sensitive setting. Even so, explanation has a defensible role when it is used to expose the features a model relied on, as Kheder et al.(24) did in a multicentre validation of an explainable gradient-boosting model for dental diagnosis, and when the explanation is designed with clinician safety in mind, as Hagiwara et al.(28)proposed for clinician-artificial intelligence interaction. The responsible position, articulated by Mihut et al.,(3) ties explainability to professional autonomy and accountability, and it is accountability rather than trust that explanation is best placed to serve. The educational dimension is relevant here, because Hudson et al.(45)and Wilson et al.(46)both reported that clinician hesitancy toward artificial intelligence is shaped by training and familiarity rather than by explanation alone, which suggests that a dental school that teaches its students how to read an explanation and when to distrust it will do more for safe adoption than a vendor that merely ships a saliency map.

The practical recommendation that follows is therefore not to abandon explanation but to change what it is asked to do, and to change it in a way that a dental clinic can implement without specialized machine-learning infrastructure, because the audit role of an explanation requires only that the model expose its attended features and its confidence, not that the clinician understand the network that produced them. A dental explanation should be presented as a check on the model, a display of the features the model weighted and the regions of the image it attended to, rather than as a justification of the model's answer, because the former invites scrutiny and the latter invites acquiescence. The distinction is the same one that separates calibrated reliance from automation bias, and it is the single most actionable lesson the evidence supports. A clinician who is shown where the model looked can compare that to where they themselves would look, and the comparison, not the explanation itself, is where the value lies. This is the position that the review's argued stance reduces to in practice, and it is one that a dental clinic could adopt today without waiting for the definitive experiment that the field still owes itself.

The limits of this review are its own, and they are stated before the conclusion. The review is narrative and non-systematic, so the selection of positions reflects a judgment about which arguments are live rather than a measurement of their prevalence. The dental evidence is thin, and much of the primary evidence is borrowed from medical specialties whose tasks differ from dental diagnosis, so the transfer of those findings to dentistry is an inference rather than a demonstration. The appraisal of evidence quality was descriptive rather than score-based, and a systematic review with a formal risk-of-bias tool would strengthen the adjudication, as the protocols of Fayaz et al.(43)and Barki et al.(44)suggest for the field as a whole. Each limit carries a remedy: a dental trust experiment that measures calibrated reliance rather than confidence, a shared standard for explanation faithfulness in dental imaging, and a systematic appraisal of the dental explainable artificial intelligence literature.

CONCLUSIONS

The claim that explainable artificial intelligence earns clinician trust in dentistry is contested rather than settled: the evidence converges on the perception that opaque models are a barrier, but it diverges on whether explanation improves accuracy and whether it calibrates or inflates trust, and each divergence traces to a concrete methodological cause, the outcome measured, the evaluation design and the sample, rather than to context. The divergence is therefore not a difference of opinion that further study will leave unresolved, but a difference of method that a single, well-designed dental experiment would settle. The argued position is that explanation is valuable in dentistry only when it is faithful, semantically meaningful in dental terms and evaluated against human decisions, and that the field has not yet produced such evidence, so the claim that explanation earns trust remains a promise to be tested rather than a result to be cited.

REFERENCES

1. Thaweesapphithak S, Thongchotchat V, Alinejad-Rokny H, Samaranayake L, Osathanon T, Porntaveetus T. Explainable Artificial Intelligence in Dentistry: A Systematic Review of Its Trust and Translation. International Dental Journal. 2026;76(4):109626. doi: 10.1016/j.identj.2026.109626
2. Al-Hammad WE, Khatatbeh MY, AlJamal G, Shatnawi MQ, Fawaeer B, Alhatamleh S, et al. Explainable artificial intelligence in dental imaging: a systematic review of interpretability and the current state of trust evidence. Front Dent Med. 2026;7:1948676. doi: 10.3389/fdmed.2026.1948676
3. Mihut T, Pantea M, Chivu RD, Popescu-Spineni DM, Cristache CM. Responsible Clinical AI in Dentistry: Trust, Professional Autonomy, and Perceptions of Accountability Across Stakeholders in Romania—A Multidisciplinary Cross-Sectional Survey. Healthcare. 2026;14(18):3003. doi: 10.3390/healthcare14183003
4. Song YJ, Song YJ, Park S, Park S, Lee YA, Lee YA, et al. Proposal of an Explainable AI (XAI)-Based Clinical Decision Support Service System for Enhancing Doctor-Patient Communication and Promoting Rational Patient Decision-Making: A Case Study in Dental Care. adr. 2026;39(3):143-173. doi: 10.15187/adr.2026.08.39.3.143
5. Vivel-Couso A, Vila-Blanco N, Carreira MJ, Bugarín-Diz A, Tomás I, Alonso-Moral JM. Trustworthy Data-driven Chronological Age Estimation from Panoramic Dental Images. Inf Syst Front. 2026. doi: 10.1007/s10796-025-10682-3
6. Can Z, Aydin E. Explainable CNN–Radiomics Fusion and Ensemble Learning for Multimodal Lesion Classification in Dental Radiographs. Diagnostics. 2025;15(16):1997. doi: 10.3390/diagnostics15161997
7. Ikerionwu C, Arungwa I, Emelogu TM, Nwabuike CE, Ukwandu E. Clinician-Centric Explainable Artificial Intelligence Framework for Medical Imaging Diagnostics: A Systematic Review. International Journal of Biomedical Imaging. 2026;2026(1):6366492. doi: 10.1155/ijbi/6366492
8. Savage CH, Sulam J, Huang CM, Yi PH. Explainable Artificial Intelligence (AI) for Medical Imaging: A Framework for Bridging the AI Trust Gap. American Journal of Roentgenology. 2026;227(2):e2634829. doi: 10.2214/ajr.26.34829
9. Chu X, Chen F, Tong M, Lu Y. Optimal presentation of decision explanations: investigating the effects of explainability and reliability on medical diagnostic performance with conversational AI. Front Psychol. 2026;17:1816652. doi: 10.3389/fpsyg.2026.1816652
10. Clement J, Ren Y, Curley S. AI Quality and Clinical Roles Impact Decision Quality in AI-Augmented Medical Decisions More than AI Explanations. Med Decis Making. 2026:0272989X261484927. doi: 10.1177/0272989x261484927
11. Brdnik S, Colakovic I, Karakatič S. Unjustified trust and satisfaction in explainable artificial intelligence: The illusion of transparency persists across user expertise, even in objectively wrong outcomes. Engineering Applications of Artificial Intelligence. 2026;182:115894. doi: 10.1016/j.engappai.2026.115894
12. Hunsicker T, Schulz A, Leist RA, Kiefer S, Boden KT, Rothaus K, et al. Efficiency Pitfalls of Explainable AI in Clinical Diagnostic and Treatment Human-AI Workflows. Hum Factors. 2026:00187208261443764. doi: 10.1177/00187208261443764
13. Okoroafor O, Ibrahim MS, Ibecheozor C, Ukatu CE, Arhin E. Explainable Artificial Intelligence, Clinician Trust and Diagnostic Accuracy in Multi-modal Medical Imaging: A Critical Narrative Review. J Adv Med Pharm Sci. 2026;28(8):111-134. doi: 10.9734/jamps/2026/v28i8884
14. Borra S, Dey N, Fong S, Sherratt RS, Shi F. Explainability and Trust in Deep Learning for Cancer Imaging: Systematic Barriers, Clinical Misalignment, and a Translational Roadmap. Cancers. 2026;18(9):1361. doi: 10.3390/cancers18091361
15. Davis SE, Salwei ME. Explainability in context: calibrating appropriate trust and reliance in artificial intelligence. J Am Med Inform Assoc. 2026;33(8):1554-1563. doi: 10.1093/jamia/ocag082
16. Haas S, Högemann M, Thomas O. Prescribing a dose of transparency: a qualitative evaluation of AI explanations with cardiovascular healthcare professionals. Front Artif Intell. 2026;9:1830201. doi: 10.3389/frai.2026.1830201
17. Nicolson A, Bradburn E, Gal Y, Papageorghiou AT, Noble JA. The human factor in explainable artificial intelligence: clinician variability in trust, reliance, and performance. npj Digit Med. 2025;8(1):658. doi: 10.1038/s41746-025-02023-0
18. Stevens AF, Stetson P. Theory of trust and acceptance of artificial intelligence technology (TrAAIT): An instrument to assess clinician trust and acceptance of artificial intelligence. Journal of Biomedical Informatics. 2023;148:104550. doi: 10.1016/j.jbi.2023.104550
19. Baskerville C, Willaime JMY, Prakash V, Ali T, Bhuva S, Ellul G, et al. Examining explainable artificial intelligence in TNM staging with PET-CT: a user-centred observation study. Radiol med. 2026;131(9):1601-1612. doi: 10.1007/s11547-026-02226-9
20. Ud Din S, Kemna R, Ket JCF, Iqbal M, Bohoudi O, Hoogendoorn M, et al. Which explainable AI methods in medical imaging are clinically impactful? A systematic literature review addressing the clinician's perspective. Front Artif Intell. 2026;9:1819422. doi: 10.3389/frai.2026.1819422
21. Van Dort BA, Engelsma T, Sneekes P, Peute L, Medlock S. Explainable AI in hospital clinical decision support systems: A scoping review of healthcare professionals' perspectives. PLOS Digit Health. 2026;5(5):e0001417. doi: 10.1371/journal.pdig.0001417
22. Pesapane F, Rotili A, Penco S, Dominelli V, Priolo F, Marinucci I, et al. Explainable Artificial Intelligence in Mammography: A Systematic Review of Methods, Evaluation Practices, and Clinical Readiness. Diagnostics. 2026;16(9):1412. doi: 10.3390/diagnostics16091412
23. Fotopoulos D, Ladakis I, Filos D, Moreno-Sánchez PA, van Gils M, Chouvarda I. Explainable AI in Cancer Imaging: Scoping Review of Methods, Modalities, and Clinical Integration. J Med Internet Res. 2026;28:e80645-e80645. doi: 10.2196/80645

24. Kheder W, Leblouba M, Rego R, Hamdoon Z. Multicentre validation and clinical interpretation of an explainable gradient-boosting model for dental-implant survival/failure prediction. Journal of Dentistry. 2025;163:106166. doi: 10.1016/j.jdent.2025.106166
25. Brown AE, Marostica CW, Hodgson NR, Martini WA. A Global and Local SHAP-Driven Interpretable Framework for Early Hospital Admission Prediction With Machine Learning at Emergency Department Triage. Cureus. 2026;18(6):e111163. doi: 10.7759/cureus.111163
26. Elsayad AM, Elsayad OA. Enhancing clinical insights in glioma grading using Bayesian Optimization and Explainable AI. Journal of Genetic Engineering and Biotechnology. 2026;24(1):100669. doi: 10.1016/j.jgeb.2026.100669
27. Huang R, Cecil J, Freedman M, Chattopadhyay S. Factors Shaping Trust and Satisfaction With AI Medical Chatbots: A Mixed Methods Vignette Survey of Caregivers Seeking Guidance on Pediatric Infectious Diseases. J Med Internet Res. 2026;28:e88126-e88126. doi: 10.2196/88126
28. Hagiwara Y, Fitch K, Trapp M. Safety design guidelines for clinician–AI interaction in computer-aided diagnosis systems using system-theoretic framework with explainability validation. Sci Rep. 2026;16(1):23561. doi: 10.1038/s41598-026-64180-w
29. Shiddik MAB. Explainable Artificial Intelligence in Healthcare: Current Landscape, Challenges, and Future Directions. Health Science Reports. 2026;9(3):e72172. doi: 10.1002/hsr2.72172
30. Zhang K, Wang D, Lin F, Xie J, Zhou W. A comprehensive review of explainable artificial intelligence in healthcare methods, evaluation, and clinical integration. iScience. 2026;29(3):115026. doi: 10.1016/j.isci.2026.115026
31. Taleban A, Sparapani R, Noffke P, Zlochiver S, Lu Q, Widlansky ME, et al. Explainable artificial intelligence in electrocardiography: A systematic review. Biomedical Signal Processing and Control. 2026;114:109325. doi: 10.1016/j.bspc.2025.109325
32. Valishirin H, S. A. FI, Brayton KA, Broschat SL. A survey of explainable AI applications in infectious disease diagnosis. BMC Infect Dis. 2026;26(1):1843. doi: 10.1186/s12879-026-14018-7
33. Srivastava K, Srivastava R. Explainable artificial intelligence (XAI) in schizophrenia: enhancing transparency in clinical decision-making. Front Syst Biol. 2026;6:1898853. doi: 10.3389/fsysb.2026.1898853
34. Aravindkumar A, Ramadoss M, Fakhruddin Ahmed SA, Sampath V, Lakshminarayanan K. Explainable AI in healthcare: a systematic review of XAI use cases in imaging, diagnostics, and rehabilitation. Front Artif Intell. 2026;9:1749527. doi: 10.3389/frai.2026.1749527
35. Paliwal G, Kumar A, Sharma KP, Bhargava D, Shrimal VM. Transformative impact of explainable artificial intelligence: bridging complexity and trust. Discov Artif Intell. 2025;5(1):51. doi: 10.1007/s44163-025-00281-1
36. Lakshmi C. Explainable Artificial Intelligence in Data Science: Enhancing Transparency and Trust in Predictive Models. IJAIML. 2026;6(13s):2153-2159. doi: 10.51483/ijaiml.3029
37. Sarma S. Explainable Artificial Intelligence in Human–AI Collaboration: Enhancing Trust and Interpretability in Decision Systems. IJAIML. 2026;6(12s):213-223. doi: 10.51483/ijaiml.6.12s.2026.213-223
38. Marchi G. Decoding the “black-box”: explainable artificial intelligence towards trustworthy advancement in respiratory medicine. Breathe. 2026;22(1):250318. doi: 10.1183/20734735.0318-2025
39. Oisakede, E. O., Daniel, R. I. A., Olawuyi, O. F., Alabi, J. O., Analikwu, C. C., Olawade, D. B. Translational gaps in Immuno-AI: From algorithmic accuracy to clinical trust. Human immunology. 2026;87(8):111774.
40. Hamrouni A, Alrumaih TNI, Ltaeif M, Hamdi M, Elmannai H, Ejbali R. Towards Explainable Artificial Intelligence in Medical Imaging: A Review of Perturbation and Backpropagation-Based Methods. JoVE. 2026;(236). doi: 10.3791/72637
41. Durak G, Aktas HE, D’Antonoli TA, Bagci U. Explainable artificial intelligence in medical imaging: how to interpret, evaluate, and use artificial intelligence explanations. dir. 2026. doi: 10.4274/dir.2026.264319
42. Mi J, Bhatt A, Kabra R, Lakkireddy D. Artificial Intelligence in Cardiac Electrophysiology Ethics, Explainability, Equity, and the Road Ahead. Indian Pacing and Electrophysiology Journal. 2026. doi: 10.1016/j.ipej.2026.10.008
43. Fayaz M, Hagsong K, Danish S, Dang LM, Sadeghi-Niaraki A, Moon H. Demystifying Artificial Intelligence: A Systematic Review of Explainable Artificial Intelligence in Medical Imaging. Sensors. 2026;26(7):2131. doi: 10.3390/s26072131
44. Barki C, Chouchen M, Sediri A, Ceylan Hİ, Rahmouni HB, Muntean RI, et al. A generative AI multi-agent framework with integrated XAI governance for cancer diagnostics: from multi-omics interpretation to lifestyle risk stratification. Front Syst Biol. 2026;6:1892418. doi: 10.3389/fsysb.2026.1892418
45. Hudson C, Phan TL, Randall M. Clinicians’ Attitudes and Perceptions on the Adoption of AI in Mental Health Care: Scoping Review. J Med Internet Res. 2026;28:e92370-e92370. doi: 10.2196/92370
46. Wilson N. Addressing Clinician-Educator Hesitancy Toward Artificial Intelligence Through a Peer-Led Instructional Design. Cureus. 2026;18(7):e113342. doi: 10.7759/cureus.113342

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Ayu Hidayat, Sari Nasution.
Writing–review and editing: Ayu Hidayat, Sari Nasution.