Clinical AI is not (yet) trustworthy-but it could be

Ali Saada, Sofia B. Dias, Ghada Alhussein, David Lyreskog, Ioannis Gerasimou, Beatriz Alvesc, Maarten de Vos, Ioannis Drivas, John Zaras, Andreas Stergioulas, Bensenousi Bensenousi, Leontios Hadjileontiadis, Christos Chatzichristos, Stelios Hadjidimitriou

Submitted to: Journal of Medical Internet Research
on: October 07, 2025

Table of Contents

Clinical AI is not (yet) trustworthy-but it could be

Ali Saada[1*]; Sofia B. Dias[2*]; Ghada Alhussein[3*]; David Lyreskog[4]; Ioannis Gerasimou[5]; Beatriz Alvesc[6]; ?aarten de Vos[7]; Ioannis Drivas[8]; John Zaras[9]; Andreas Stergioulas[10]; Bensenousi Bensenousi[1]; Leontios Hadjileontiadis[5, 3*] Prof Dr; Christos Chatzichristos[7*]; Stelios Hadjidimitriou[5*]

[1] AINIGMA Technologies Leuven BE
[2]Faculdade de Motricidade Humana Universidade de Lisboa Centro Interdisciplinar de Estudo da Performance Humana Lisbon PT
[3]Department of Biomedical Engineering and Biotechnology College of Medicine and Health Sciences Khalifa University of Science and Technology Abu Dhabi AE
[4]NEUROSEC Department of Psychiatry University of Oxford Oxford GB
[5]Dept. of Electrical and Computer Engineering School of Engineering Aristotle University of Thessaloniki Thessaloniki GR
[6] Faculdade de Motricidade Humana University of Lisbon Lisbon PT
[7]STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics Department of Electrical Engineering KU Leuven Leuven BE
[8] Diadikasia Business Consulting Symvouloi Epicheiriseon AE Athens GR
[9] Squaredev Brussels BE
[10] Information Technologies Institute Centre for Research and Technology Hellas Thessaloniki GR
[*]these authors contributed equally

Corresponding Author:
Leontios Hadjileontiadis Prof Dr
Dept. of Electrical and Computer Engineering
School of Engineering
Aristotle University of Thessaloniki
Univ. Campus, D Building, 6th floor, Of#26
Thessaloniki
GR

Abstract

The shift toward trustworthy artificial intelligence (AI) in healthcare marks a pivotal transformation. Traditionally, clinical AI systems have lacked dynamic trust integration across their lifecycle. With structured governance frameworks, AI in healthcare is evolving—ushering in a new era of trust-enabling technologies. In this Viewpoint, we present a framework grounded in the Assessment List for Trustworthy Artificial Intelligence (ALTAI) and applied within the Horizon Europe AI-PROGNOSIS project to embed ethical, technical, and regulatory safeguards across the AI lifecycle. By surfacing implementation tensions and integrating normative, technical, and regulatory safeguards, we outline a replicable path for building adaptive, trust-enabling infrastructures in clinical practice, demonstrating that while clinical AI is not yet trustworthy, structured, lifecycle-oriented governance makes it possible.

(JMIR Preprints 07/10/2025:85433)
DOI: https://doi.org/10.2196/preprints.85433

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

Viewpoint

Clinical AI is not (yet) trustworthy-but it could be

Ali Saad[a,†], Sofia B. Dias[b,†], Ghada Alhussein[c,d†], David Lyreskog[e], Ioannis Gerasimou[f], Beatriz Alves[c], Maarten de Vos[g], Ioannis Drivas[h], John Zaras[i], Andreas Stergioulas[j], Alex Bensenousi[a], Leontios J. Hadjileontiadis[d,f,✉], Christos Chatzichristos[g,*✉], and Stelios Hadjidimitriou[f,*✉], on behalf of the AI-PROGNOSIS Consortium‡*

[a]*AINIGMA Technologies, Leuven, Belgium*
[b]*Interdisciplinary Centre for the Study of Human Performance (CIPER), Faculdade de Motricidade Humana, Universidade de Lisboa, Lisbon, Portugal*
[c]*Faculdade de Motricidade Humana, Universidade de Lisboa, Lisbon, Portugal*
[d]*Department of Biomedical Engineering and Biotechnology, Khalifa University, Abu Dhabi, UAE*
[e]*NEUROSEC, Department of Psychiatry, University of Oxford, Oxford, UK*
[f]*Department of Electrical Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece*
[g]*Department of Electrical Engineering, STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics, KU Leuven, Leuven, Belgium*
[h]*Diadikasia Business Consulting Symvouloi Epicheiriseon AE, Athens, Greece.*
[i]*Squaredev, Brussels, Belgium*
[j]*Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece*

†These authors shared first authorship
**These authors shared last authorship*
‡Members are listed at the end of the article

✉*Corresponding authors: Prof. Leontios Hadjileontiadis; Email: leontios@auth.gr, Dr Christos Chatzichristos; Email: christos.chatzichristos@kuleuven.be, Dr Stelios Hadjidimitriou; Email: shadjidim@ece.auth.gr*

Department of Electrical and Computer Engineering
Aristotle University of Thessaloniki
GR 54124 Thessaloniki, Greece

Abstract

The shift toward trustworthy artificial intelligence (AI) in healthcare marks a pivotal transformation. Traditionally, clinical AI systems have lacked dynamic trust integration across their lifecycle. With structured governance frameworks, AI in healthcare is evolving—ushering in a new era of trust-enabling technologies. In this *Viewpoint*, we present a framework grounded in the Assessment List for Trustworthy Artificial Intelligence (ALTAI) and applied within the Horizon Europe AI-PROGNOSIS project to embed ethical, technical, and regulatory safeguards across the AI lifecycle. By surfacing implementation tensions and integrating normative, technical, and regulatory safeguards, we outline a replicable path for building adaptive, trust-enabling infrastructures in clinical practice, demonstrating that while clinical AI is not yet trustworthy, structured, lifecycle-oriented governance makes it possible.

Keywords: Trustworthy artificial intelligence (AI); Clinical AI; ALTAI Framework; AI-PROGNOSIS European research initiative; Lifecycle safeguards; Ethical AI integration

Introduction

Artificial intelligence (AI) continues to expand its footprint in healthcare, with the global AI healthcare market projected to grow from USD 20.9 billion in 2024 to USD 148.4 billion by 2029, reflecting a compound annual growth rate of 48.1[1]. While this technological momentum has led to significant gains in diagnostic accuracy, prognostic modeling, and treatment optimization, the integration of AI systems into routine healthcare remains cautious and uneven. This reticence, while partially attributable to regulatory inertia and data access limitations, is fundamentally rooted in a deeper concern: the perceived trustworthiness of AI-driven systems [2,3]. This is in fact a critical challenge, as performance alone is insufficient without ensuring reliability, ethical alignment, and public trust [4], especially as healthcare AI systems transition from proof-of-concept to clinical deployment.

Trust in clinical AI transcends conventional performance metrics. It is not reducible to algorithmic accuracy or validation statistics alone, but rather represents a composite property that encompasses transparency, interpretability, accountability, and alignment with both clinical values and ethical principles [5,6]. Recent works have highlighted how the implementation of AI in healthcare is increasingly shaped not only by technical feasibility, but by the complex interplay of governance, institutional norms, and frontline practices [7,8]. Trustworthiness in AI is increasingly recognized as a multidimensional construct that encompasses not only compliance with technical benchmarks, but also alignment with ethical principles and regulatory standards [9]. Such an alignment must extend beyond end-product validation to permeate the entire lifecycle of system development. Procedural approaches, those that embed trust-oriented safeguards from design to deployment, are essential to achieving this [10]. From the Viewpoint of end-users-including clinicians, patients, and institutions-trust is not simply earned by retrospective audit or certification; instead, it is cultivated over time, shaped by system behavior, user experience, and organizational context [11].

Despite the proliferation of frameworks and normative guidance, ranging from

high-level ethical principles to emerging regulatory instruments, a persistent implementation gap remains [11,12]. Existing instruments frequently emphasize outcomes rather than mechanisms; they evaluate trust *post hoc*, rather than embedding it procedurally throughout the AI lifecycle. This disconnect highlights the absence of a pragmatic scaffolding to guide trustworthy AI design and deployment within high-stakes environments, such as healthcare.

This Viewpoint advances a procedural approach to trustworthiness in clinical AI, drawing upon the Assessment List for Trustworthy Artificial Intelligence (ALTAI) developed by the High-Level Expert Group on AI [13], as a representative tool for operationalizing ethical and regulatory principles. Unlike purely aspirational codes, the ALTAI framework offers a practical, step-by-step checklist that can be directly integrated into project workflows. It delivers concrete guidance throughout the entire AI lifecycle, including design specification, data governance, model development, evaluation, and deployment, and includes defined metrics and procedures to promote compliance and transparency [13]. To contextualize this approach, we examine the AI-PROGNOSIS project (https://www.ai-prognosis.eu/), a European initiative focused on the development of predictive models for Parkinson's Disease (PD), as a case study. The project aims to deliver a suite of AI-driven tools for risk assessment, prognosis, and disease management. Specifically, the system integrates data from multiple sources, including patient records and clinical databases, to (1) estimate individual PD risk through probabilistic scoring, (2) predict changes in PD progression over time, and (3) assess expected responses to medication, including potential side effects. Rather than positioning the AI-PROGNOSIS project itself as the contribution, it serves to illustrate how a procedural framework can shape real-world system design.

By anchoring trustworthiness in procedural steps rather than retrospective assessments, this Viewpoint contributes to a growing body of literature calling for actionable strategies to embed ethical and regulatory principles into real-world AI systems [7,8]. In the sections that follow, we articulate the conceptual rationale for procedural trust, outline the ALTAI framework, illustrate its instantiation across the AI development continuum via the AI-PROGNOSIS case study, and reflect on the practical, ethical, and sociotechnical tensions encountered during implementation in healthcare innovation.

From principles to procedure: why trust in AI needs a blueprint

The proliferation of ethical guidelines for AI has revealed a growing consensus: trust is essential for the responsible deployment of AI in healthcare. Yet, despite the emergence of high-level principles, such as fairness, transparency, and accountability, there remains a persistent gap between normative aspirations and practical implementation [14,15]. This disconnect has prompted calls for procedural frameworks that can translate abstract values into actionable design and governance strategies [16]. This approach has several benefits to traditional principle-based frameworks of trust, which typically rely on specific targets and retrospective evaluations.

Nevertheless, trust in AI is neither a monolithic concept nor an intrinsic property of the system; rather, it is a system-level outcome shaped by dynamic interactions among users, institutions, and broader sociotechnical environments [17]. It emerges from the interplay of interrelated technical and ethical

dimensions, such as explainability, robustness, and fairness, none of which are sufficient in isolation. Cultivating trust therefore entails iterative, lifecycle-spanning processes that embed normative safeguards into the design, development, deployment, and governance of AI systems [18,19].

In alignment with the ALTAI framework [13] and ISO/IEC TS 5723:2022 [20], we identify eight core dimensions, i.e., robustness, generalization, explainability, accountability, transparency, reproducibility, fairness, and privacy, as foundational to Trustworthy AI in healthcare. These dimensions reflect a synthesis of normative principles and practical requirements for clinical-grade AI systems. In particular, *robustness* refers to a system's capacity to perform reliably under uncertainty—such as noisy inputs, adversarial perturbations, or incomplete records—without significant loss of function [9]. However, robustness must coexist with usability and interpretability, particularly in clinical environments. Closely related is *generalization*, the model's ability to extrapolate to unseen data, which remains a fundamental challenge given the risk of underfitting or overfitting, especially in small or biased datasets [18,21,22].

Explainability is a context-sensitive construct that varies across stakeholders—clinicians, patients, regulators, and developers [19]. It may be achieved through post hoc methods (e.g. SHAP, LIME) or interpretable models, each with trade-offs in fidelity and scalability [23]. *Accountability* demands clear traceability of decisions to responsible entities [24], while *transparency*, its enabling counterpart, requires open disclosure of model purpose, data provenance, and performance characteristics [25]. Achieving transparency often involves managing organizational or proprietary constraints. *Reproducibility*, a pillar of scientific integrity, remains elusive in machine learning due to non-determinism in training processes and hardware dependencies [26]. *Fairness*, arguably the most socially charged dimension, seeks to mitigate bias introduced during data collection, model design, and deployment [9,27]. Technical responses span pre-, in-, and post-processing interventions and must be informed by socioethical theories of discrimination and equity [21,27,28]. Finally, *privacy* safeguards not only identifiers but also individual autonomy over data use [12]. While techniques like differential privacy, de-identification, and data minimization offer protection, they may constrain model expressiveness or transparency [29]. These trade-offs underscore the necessity of procedural frameworks that treat trust not as a checklist but as a dynamic property shaped by ongoing design, validation, and governance.

While often treated as discrete targets, the aforementioned dimensions are deeply interwoven and must be addressed through integrated, context-sensitive design strategies across the AI lifecycle [22]. Attempts to enhance one dimension, such as increasing transparency, can inadvertently compromise another, such as protecting proprietary data or patient privacy. This interplay reinforces the need for procedural frameworks that consider trustworthiness as a system-wide property, developed iteratively and contextually throughout the AI lifecycle.

ALTAI as a procedural anchor

While core dimensions of Trustworthy AI, such as robustness, fairness, and transparency, provide conceptual structure, their realization in clinical settings

requires systematic, context-sensitive implementation. From the Asilomar AI Principles [30] and the Montreal Declaration [31], to institutional efforts like AI4People [6] and the OECD guidelines [32], much of this work has emphasized normative commitments—fairness, accountability, transparency, and safety. National strategies, including those from China [33], the United Kingdom [34], the United States [35], and the European Union [36], have begun translating these principles into policy and regulation. In the healthcare domain, oversight by bodies such as the US FDA [35] and NIST [37] adds additional complexity, particularly for high-risk systems.

Among these efforts, the EU's Assessment List for Trustworthy Artificial Intelligence (ALTAI) [38], developed by the High-Level Expert Group on Artificial Intelligence (HLEG-AI) in 2020 [32], stands out as a concrete procedural framework for embedding trust across the AI development lifecycle. Unlike principle-driven charters, ALTAI codifies seven actionable requirements: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity and fairness, societal well-being, and accountability. These dimensions are designed not as abstract endpoints, but as iterative checkpoints, aligning design practices with ethical and regulatory imperatives at each stage of AI development. Its web-based tool supports structured self-assessment and generates visual diagnostics, such as radar plots, summarizing strengths and deficiencies, thus enabling continuous monitoring, recommended next steps, and targeted refinement.

Albeit scoring details remain opaque, ALTAI represents one of the most widely adopted instruments for proceduralizing trust in AI workflows. To assess its translational value in applied clinical research, we adapted ALTAI within the AI-PROGNOSIS project. This adaptation aligned technical design efforts with structured trust requirements, providing a foundation for identifying risk points, embedding ethical safeguards, and supporting internal reflection among development teams. In the following section, we present this case study as an applied instantiation of ALTAI, illustrating how trust-oriented governance can be realized through procedural integration.

Operationalizing procedural trust: the AI-PROGNOSIS case study

To evaluate the ALTAI procedural trust framework in clinical AI, we conducted a structured assessment within the AI-PROGNOSIS project, that aims to generate individualized PD risk scores, PD progression forecasts, and medication response estimates, using machine learning techniques applied to multimodal health data.

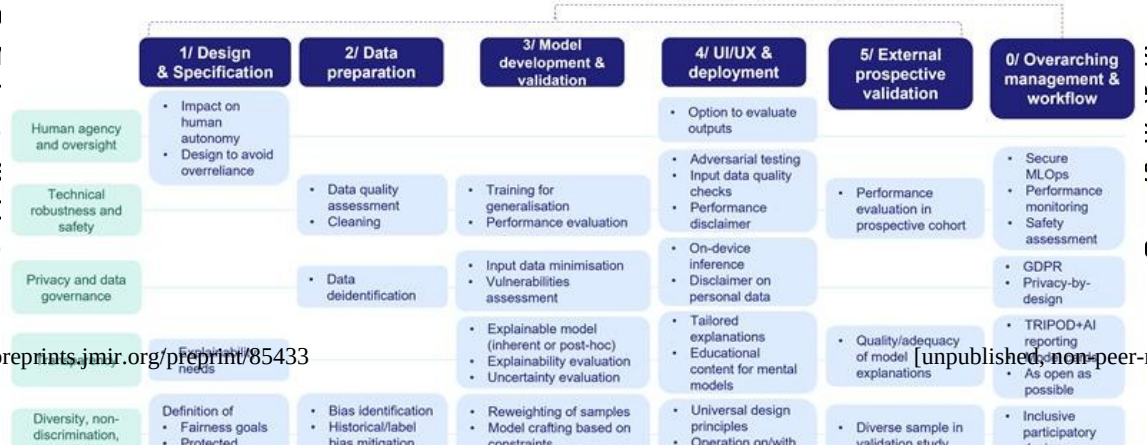
More

valuate
ondents
itutions
um. The
age of
erience.
pipeline

7 ALTAI
mmental
impact, using a three-point scale (low, medium, high). The main aim was to identify which components were considered most critical for a healthcare-focused AI system in development. Overall, survey responses revealed a consistent emphasis on technical and ethical priorities.

Figure 1 illustrates the average importance ratings provided by AI experts across the seven ALTAI requirements. More specifically, subgroups related to accuracy, reliability, fallback planning, data privacy, bias mitigation, and stakeholder participation received the highest average importance ratings. In co such as
er ss critical,
re jection of
th e granular
vi subgroup.
Ap alongside
th o support

Figure 2. Trustworthy AI components embedded in the AI-PROGNOSIS framework.
The graphical representation summarizes how dimension-specific explainability, user interface (UI), user experience (UX), and accountability are operationalized within the AI-PROGNOSIS project, supporting trust by design. AI: Artificial Intelligence.



integration across the AI-PROGNOSIS development cycle, ALTAI's seven requirements were mapped to six core lifecycle stages, namely: design and specification, data preparation, model development and validation, interface design and deployment, external validation, and cross-cutting governance and workflow (see Figure 2). This mapping, also presented in Appendix Table S1, aligns with HLEG-AI's guidance on adapting trustworthiness frameworks to specific system contexts. It enabled the identification of stage-specific trust touchpoints and informed the implementation of safeguards, such as adversarial testing, explainability benchmarking, and privacy-preserving data handling.

Beyond diagnostic value, the ALTAI adaptation fostered internal reflection among development teams, prompting early-stage deliberation on fallback mechanisms, user interface accessibility, and the ethical implications of probabilistic risk scoring. These insights informed both system architecture and stakeholder engagement strategies, reinforcing the role of procedural frameworks as catalysts for trust-aware design. Collectively, these findings emphasize ALTAI's function not merely as an evaluative instrument, but as a formative scaffold for embedding trust-oriented design logic throughout the development pipeline, as well. This case study exemplifies how procedural frameworks can facilitate the operationalization of ethical and regulatory objectives in clinical AI, with technical teams serving as key intermediaries in translating abstract principles into implementable system architecture.

Tensions in pursuit of Trustworthy AI

While procedural frameworks such as ALTAI offer structured guidance for embedding trust into AI development, their translation into clinical practice reveals persistent tensions. These tensions are not solely technical but emerge from the entanglement of ethical, organizational, and regulatory constraints with real-world implementation dynamics. Trust, correspondingly, is not a static attribute in this context, but must be seen as a negotiated, adaptive property shaped and actively maintained across the AI lifecycle by design choices, deployment decisions, and end-user interactions[16,30]. Staking out paths to navigate these tensions should therefore be a procedural undertaking, allowing dynamic adaptation on a case-to-case basis, yet offering a guiding rail. Below, we outline key decision-points, and how they were navigated in the AI-PROGNOSIS project, using this approach.

Engineering clarity under complexity

Design-time interventions within AI-PROGNOSIS emphasized risk mitigation through adversarial robustness, constrained model behavior, and transparency mechanisms. Structured data validation pipelines were established to ensure completeness, consistency, and semantic fidelity of clinical inputs. Early-stage vulnerability assessments leveraged tools such as the Adversarial Robustness Toolbox [39] and CleverHans [40] to stress-test model responses under plausible perturbations. These safeguards were necessary but not sufficient, given that trust is also shaped by user perceptions, system intelligibility, and the sociolegal context of deployment.

Explainability was pursued through SHAP and LIME visualizations [41], integrated into intuitive user experience/user interface (UX/UI) elements, co-developed through co-design sessions with patients, clinicians, and human-

computer interaction specialists. The desired output included layout consistency, content hierarchy, accessibility, and performance optimization that can ensure usability across user populations. Additionally, design principles, such as Google's Material Design or Apple's Human Interface Guidelines offered baseline frameworks for further validation through iterative testing with diverse users. These interfaces incorporated risk disclaimers, scenario-specific warnings, and model output rationales, elements shown to foster situational awareness and calibrate expectations [42]. However, tension emerged between model interpretability and predictive fidelity: simpler, more explainable models sometimes underperformed in capturing longitudinal patterns, while high-dimensional neural architecture offered superior accuracy at the cost of intelligibility [43,44].

Transparency obligations are also intersected with institutional and commercial constraints. Open disclosure of model behavior, data lineage, and source code encountered resistance when proprietary IP, reputational risk, or liability exposure were perceived. These experiences echo concerns documented in broader AI governance literature, where explainability is seen as a boundary object, interpreted differently by legal experts, regulators, engineers, and lay users [16,19,45]. Trust, therefore, cannot rely solely on *post hoc* visualization tools or interface overlays; it must be cultivated through continuous interaction between technical artifacts and epistemic communities.

Generalization versus representativeness

Achieving generalization in clinical AI extends beyond algorithmic optimization. It requires validating performance across subpopulations, healthcare settings, and temporal shifts, domains where real-world complexity and structural inequities surface. Within AI-PROGNOSIS, data were curated to reflect heterogeneity across age, sex, and disease severity, with feature reviews conducted alongside clinicians to de-risk unintentional bias proxies.

Still, key fairness checks were constrained by unavailable or restricted variables. Under General Data Protection Regulation (GDPR) [46] and ethical review protocols, collection of race, ethnicity, and socioeconomic indicators was either prohibited or discouraged, limiting the granularity of bias auditing [9,27,28]. These limitations underscored the tension between privacy-preserving practice and equity-informed auditing. Without disaggregated data, even well-calibrated models can systematically underperform or generate disparate outcomes for minority groups [47]. Moreover, generalizability was not static. Drift monitoring using Evidently AI [36] enables temporal performance tracking but requires careful configuration to avoid false alarms or blind spots. Participatory workshops were convened with diverse stakeholders to co-define performance thresholds, identify context-specific harms, and design explanation strategies tailored to each stakeholder type [44,48]. In the same vein, external validation ensures that AI systems generalize beyond development data. Explainable AI (xAI) plays a central role during external validation. Explanation strategies should be adapted to different user groups, from clinicians to patients to engineers. At the same time, co-creation workshops or equivalent co-creation processes (e.g., innovation jams, living labs, open innovation platforms, lead user collaborations, and crowdsourced design challenges), are valuable for defining these strategies and ensuring a shared understanding of outputs.

These findings suggest that generalization must be reframed: not only as an

empirical measure of cross-sample performance, but as a sociotechnical process of aligning predictive behavior with real-world variance, regulatory constraints, and stakeholder expectations. Fairness, interpretability, and robustness cannot be optimized independently; they must be co-engineered through continuous iteration and value-sensitive design [15,49].

Institutional scaffolding for sustainable trust

Trust, to be enduring, must outlive model deployment. Within AI-PROGNOSIS, governance mechanisms were embedded across the system lifecycle using MLOps pipelines [50] managed via MLflow [51]. These supported reproducibility, lineage tracking, and automated logging of model outputs for compliance and audit purposes. Model documentation, following TRIPOD and Datasheets for Datasets [52] guidelines, facilitates reproducibility and informs external reviewers about data provenance, development conditions, and known limitations. Ethical and legal oversight structures were supported through dynamic data governance plans housed in European Open Science Cloud's ARGOS [53], which offers version-controlled templates for privacy, security, and access control compliance. An internal ethics board, comprising technical, legal, and clinical representatives, was tasked with monitoring value drift, assessing updates to explainability outputs, and coordinating stakeholder feedback loops.

Sustainability efforts address both environmental impact (e.g., energy cost of training pipelines) and downstream clinical implications (e.g., deskilling risks or overdependence on AI predictions) [15]. These are increasingly salient concerns as healthcare systems adopt AI at scale and require not only functional models, but systems that preserve professional autonomy and adapt to evolving sociopolitical contexts [15,49]. Moreover, regulatory compliance, particularly with the GDPR [46], is non-negotiable. Privacy-by-design principles should guide system architecture. Data management teams must ensure ethical oversight, secure processing, and adherence to data-sharing agreements.

Crucially, trust must be institutionally maintained. This requires aligning development workflows with adaptive governance structures capable of incorporating feedback, absorbing policy shifts, and ensuring ethical continuity over time. Static checklists are ill-suited for this role; procedural frameworks, on the other hand, can (and must) evolve into organizational capabilities, rooted in accountability, reflexivity, and stakeholder engagement.

Outlook: building living systems of trust

This Viewpoint has outlined a procedural approach to embedding trust in clinical AI, grounded in the ALTAI framework [38] and instantiated through the AI-PROGNOSIS project. By integrating ethical, technical, and regulatory safeguards across the AI lifecycle, we have demonstrated how trust can be operationalized not only as a design goal but as a dynamic property of clinical AI systems.

Yet, as the field matures, it is increasingly clear that procedural scaffolding alone has its limitations. Trustworthiness must be sustained through adaptive governance, capable of responding to evolving risks, shifting stakeholder expectations, and emerging regulatory mandates. In this regard, the newly introduced European Artificial Intelligence Act (AI Act) [54] represents a pivotal inflection point. Entering into force in August 2024, the AI Act introduces a harmonized, risk-based legal framework for AI across the EU, with specific obligations for high-risk systems, including those deployed in healthcare [54].

These include requirements for transparency, human oversight, robustness, and post-market monitoring, many of which align with ALTAI's procedural ethos but now carry legal enforceability.

The AI-PROGNOSIS framework is being continuously adapted to anticipate and respond to these and other regulatory and policy developments. Specifically, future iterations will (1) integrate AI Act compliance checkpoints into development workflows, (2) expand stakeholder engagement to include legal and regulatory experts, and (3) establish mechanisms for continuous post-deployment monitoring and redress. These steps reflect the broader approach, shifting away from principle-based ethics towards procedural trust and institutionalized accountability, where trust is not only designed and cultivated, but governed.

Looking ahead, we argue that Trustworthy AI in healthcare must be conceptualized as a living system, one that evolves through iterative feedback, interdisciplinary and diverse collaboration, and regulatory responsiveness. This requires moving beyond static checklists toward reflexive infrastructures that embed ethical deliberation, stakeholder negotiation, and lifecycle oversight into the core of AI development. As the regulatory landscape crystallizes and clinical adoption accelerates, such infrastructures will be essential to ensure that AI systems remain not only performant, but aligned with the values, rights, and expectations of the societies they serve.

Acknowledgements

This study receives funding from the European Union under Grant Agreement No. 101080581 (AI-PROGNOSIS). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency. Neither the European Union nor the European Health and Digital Executive Agency can be held responsible for them. Finally, we thank all members of the AI-PROGNOSIS Consortium. Moreover, LJH acknowledges support from Khalifa University of Science and Technology, Abu Dhabi, UAE, Provost's Office Grant. The funding sources had no role in study design, data collection, data analysis, data interpretation, the writing of the manuscript or the decision to submit it for publication.

AI-PROGNOSIS Consortium

Leontios J. Hadjileontiadis[1], Stelios Hadjidimitriou[1], Vasileios Charisis[1], Ioannis Gerasimou[1], Charalampos Sotirakis[1], Apostolos Moustaklis[1], Thanasis Kakasis[2], Ali Saad[2], Nikos Melanitis[2], Alex Bensenousi[2], Angeliki Zarifi[2], Despina Anastasopoulos[3], Dorine Matzakou[3], Theodora Brisimi[3], Marilena Damkali[3], Petros-Sozon Dimitrakopoulos[3], Alexandros Miouglou[3], Ioannis Drivas[4], Sotirios Michagiannis[4], Therese Scott Duncan[5], Jamie Luckhaus[5], Anna Clareborn[5], Sara Riggare[5], Olga Sanchez Solino[6], Amel Drif[7], Anna Rybicka[7], Natalia Del Campo[7], Margherita Fabbri[7], Olivier Rascol[7], Sofia B. Dias[8], Ghada Alhussein[8], Beatriz Alves[8], Filomena Carnide[8], Nikos Grammalidis[9], Kosmas Dimitropoulos[9], Andreas Stergioulas[9], Theocharis Chatzis[9], Nikola Goetz[10], Niloofar Tavakoli[10], Helene Huts[11], Maarten De Vos[11], Christos Chatzichristos[11], Thomas Strypsteen[11], Fan Wang[11], Aldona Niemiro Sznajder[11], Georgios Roussis[11], Elissavet Zogopoulou[12], Charis Giaralis[12], John Zaras[12], Christos Vasilakis[12], Björn Falkenburger[13], Nils Schnalke[13], Tim Feige[13], Kristina Leipuviene[14], Eleni Zamba-Papanicolaou[15], Kyriaki Michailidou[15], Christiana Christodoulou[15], Paraskevi Chairta[15], Kyproula Christodoulou[15], David Lyreskog[16], Maria-Luisa Almarcha-Menargues[17], and Monica Kurtis Urra[17].

[1]Aristotle University of Thessaloniki, Greece. [2]Ainigma Technologies, Greece. [3]Netcompany-Intrasoft, Greece. [4]Diadikasia Business Consulting Symvouloi Epicheiriseon, Greece. [5]Uppsala University, Sweden. [6]Abbvie Deutschland GmbH & Co. KG, Spain. [7]Centre Hospitalier Universitare de Toulouse, France. [8]Faculdade de Motricidade Humana, University of Lisbon, Portugal. [9]Centre for Research and Technology Hellas, Greece. [10]Neurotransdata GmbH, Germany. [11]Katholieke Universiteit Leuven, Belgium. [12]SquareDev, Greece. [13]Technische Universitaet Dresden, Germany. [14]Smartsol SIA,

Lithuania. [15]Kypriako Idryma Erevnon Gia Ti Myiki Distrofia, Cyprus. [16]University of Oxford, UK. [17]Movement Disorders Unit, Neurology Department, Hospital Ruber Internacional, Madrid, Spain.

Data Availability

The datasets generated and analyzed during this study are not publicly available, but they can be accessed upon reasonable request from the corresponding author.

Authors' Contributions

AS, SH, IG, DL, MV, ID, JZ, AB, LJH, CC developed the concept of the manuscript and AS prepared the first draft of the manuscript. SBD, GA, DL, SH, CC, LJH contributed to the direction of the manuscript and provided ongoing feedback, reviewing the manuscript critically for important intellectual content. All authors edited and reviewed the final manuscript. All authors reviewed and accepted the final version of the paper and agreed to the decision to submit it. All authors reviewed and accepted the final version of the paper, were not precluded from accessing data in the study, and they accepted responsibility to submit for publication. LJH and SH have accessed and verified the data.

Conflicts of Interest

None declared.

Multimedia Appendix

Average importance score by ALTAI subgroup and full set of ALTAI checklist responses alongside the corresponding system-generated recommendations.

References

1. Market Research Future. AI in healthcare market statistics forecast to 2029. 2024. URL: https://www.marketresearchfuture.com/reports/ai-in-healthcare-market-6980 [accessed 2025-10-07]
2. Nong P, Platt J. Patients' trust in health systems to use artificial intelligence. *JAMA Netw Open* 2025;8:e2460628. [doi: 10.1001/jamanetworkopen.2024.60628]
3. Tucci V, Saary J, Doyle TE. Factors influencing trust in medical artificial intelligence for healthcare professionals: a narrative review. *J Med Artif Intell* 2022;5:e000001. [doi: 10.21037/jmai-22-1]
4. Khalighi S, et al. Artificial intelligence in neuro-oncology: advances and challenges in brain tumor diagnosis, prognosis, and precision treatment. *NPJ Precis Oncol* 2024;8:1–12. [doi: 10.1038/s41698-024-00345-2]
5. Nickel PJ. Trust in medical artificial intelligence: a discretionary account. *Ethics Inf Technol* 2022;24:e000002. [doi: 10.1007/s10676-022-09613-9]
6. Floridi L, et al. AI4People—An ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Mach* 2018;28:689–707. [doi: 10.1007/s11023-018-9482-5]
7. Carboni C, Brightwell C, Halpern O, Freyer O, Gilbert S. Reconciling security and care in digital medicine. *NPJ Digit Med* 2025;8:1–5. [doi: 10.1038/s41746-025-00321-9]
8. Bodnari A, Travis J. Scaling enterprise AI in healthcare: the role of governance in risk mitigation frameworks. *NPJ Digit Med* 2025;8:1–4. [doi: 10.1038/s41746-025-00318-4]
9. Tran M, et al. Situating governance and regulatory concerns for generative artificial intelligence and large language models in medical education. *NPJ Digit Med* 2025;8:1–10. [doi: 10.1038/s41746-025-00325-5]
10. Reinhardt K. Trust and trustworthiness in AI ethics. *AI Ethics* 2023;3:735–744. [doi: 10.1007/s43681-023-00242-1]
11. Kerasidou C, Kerasidou A, Buscher M, Wilkinson S. Before and beyond trust: reliance in medical AI. *J Med Ethics* 2022;48:852–856. [doi: 10.1136/medethics-2021-107791]
12. Durán JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *J Med Ethics* 2021;47:e000003. [doi: 10.1136/medethics-2020-106820]

13. European Commission. High-level expert group on artificial intelligence. URL: https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai [accessed 2025-10-07]
14. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell* 2019;1:389–399. [doi: 10.1038/s42256-019-0088-2]
15. Mittelstadt B. Principles alone cannot guarantee ethical AI. *Nat Mach Intell* 2019;1:e000004. [doi: 10.1038/s42256-019-0114-4]
16. Dignum V. Responsible artificial intelligence: how to develop and use AI in a responsible way. *Ariz State Law J* 2019;51:501–507. URL: https://arizonastatelawjournal.org/responsible-ai-dignum [accessed 2025-10-07]
17. Hamon R, Junklewitz H. Robustness and explainability of artificial intelligence: from technical to policy solutions. *EUR 30040 EN*, Publications Office of the European Union; 2020. [doi: 10.2760/192653]
18. Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning (still) requires rethinking generalization. *Commun ACM* 2021;64:107–115. [doi: 10.1145/3446776]
19. Papagni G, de Pagter J, Zafari S, Filzmoser M, Koeszegi ST. Artificial agents' explainability to support trust: considerations on timing and context. *AI Soc* 2023;38:947–960. [doi: 10.1007/s00146-022-01415-4]
20. International Organization for Standardization. ISO/IEC TS 5723:2022—Trustworthiness: vocabulary. URL: https://www.iso.org/standard/81608.html [accessed 2025-10-07]
21. Chen P, Wu L, Wang L. AI fairness in data management and analytics: a review on challenges, methodologies and applications. *Appl Sci* 2023;13:10258. [doi: 10.3390/app131710258]
22. Li B, et al. Trustworthy AI: from principles to practices. *ACM Comput Surv* 2023;55:1–46. [doi: 10.1145/3507950]
23. Retzlaff CO, et al. Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists. *Cogn Syst Res* 2024;86:101243. [doi: 10.1016/j.cogsys.2023.101243]
24. Cui L, Qu Y, Gao L, Xie G, Yu S. Detecting false data attacks using machine learning techniques in smart grid: a survey. *J Netw Comput Appl* 2020;170:e000005. [doi: 10.1016/j.jnca.2020.102795]
25. Geisler S, et al. Knowledge-driven data ecosystems toward data transparency. *J Data Inf Qual* 2022;14:1–12. [doi: 10.1145/3491141]
26. Cockburn A, Dragicevic P, Besançon L, Gutwin C. Threats of a replication crisis in empirical computer science. *Commun ACM* 2020;63:70–79. [doi: 10.1145/3376898]
27. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Comput Surv* 2021;54:1–35. [doi: 10.1145/3457607]
28. Binns R. Fairness in machine learning: lessons from political philosophy. *Proc Mach Learn Res* 2018;81:149–159. URL: https://proceedings.mlr.press/v81/binns18a.html [accessed 2025-10-07]
29. Agarwal S. Trade-offs between fairness and privacy in machine learning. *IJCAI Workshop AI for Social Good* 2021. URL: https://www.ijcai.org/proceedings/2021/workshop/ai4sg [accessed 2025-10-07]
30. Future of Life Institute. Asilomar AI Principles. URL: https://futureoflife.org/ai-principles [accessed 2025-10-07]
31. Ménissier T. A "Machiavellian moment" for artificial intelligence? The Montreal Declaration for the responsible development of AI. *Raisons Polit* 2020;77:67–81. [doi: 10.3917/rai.077.0067]
32. European Commission. High-level expert group on artificial intelligence. URL: https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai [accessed 2025-10-07]
33. Governance Principles for the New Generation Artificial Intelligence. URL: https://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html [accessed 2025-10-07]
34. House of Lords Artificial Intelligence Committee. AI in the UK: ready, willing and able? URL: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm [accessed 2025-10-07]
35. United States Food & Drug Administration. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD). 2020. URL: https://www.fda.gov/media/122535/download [accessed 2025-10-07]
36. European Parliament. EU AI Act: first regulation on artificial intelligence. URL: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-

regulation-on-artificial-intelligence [accessed 2025-10-07]
37. Tabassi E. Artificial Intelligence Risk Management Framework (AI RMF 1.0). 2023. [doi: 10.6028/NIST.AI.100-1]
38. Rajamäki J, et al. ALTAI tool for assessing AI-based technologies: lessons learned and recommendations from SHAPES pilots. *Healthcare* 2023;11:1454. [doi: 10.3390/healthcare11091454]
39. Nicolae MI, et al. Adversarial robustness toolbox v1.0.0. 2018. URL: https://github.com/Trusted-AI/adversarial-robustness-toolbox [accessed 2025-10-07]
40. Goodfellow I, Papernot N, McDaniel P. Cleverhans v0.1: an adversarial machine learning library. 2016. [doi: 10.48550/arXiv.1610.00768]
41. Band S, et al. Application of explainable artificial intelligence in medical health: a systematic review of interpretability methods. *Inform Med Unlocked* 2023;40:101286. [doi: 10.1016/j.imu.2023.101286]
42. Jawaheer G, Weller P, Kostkova P. Modeling user preferences in recommender systems: a classification framework for explicit and implicit user feedback. *ACM Trans Interact Intell Syst* 2014;4:1–26. [doi: 10.1145/2559982]
43. Watson DS, et al. Clinical applications of machine learning algorithms: beyond the black box. *BMJ* 2019;364:l886. [doi: 10.1136/bmj.l886]
44. Nazar M, Alam MM, Yafi E, Su'Ud MM. A systematic review of human-computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques. *IEEE Access* 2021;9:153316–153348. [doi: 10.1109/ACCESS.2021.3127985]
45. Lipton ZC. The mythos of model interpretability. *Commun ACM* 2018;61:36–43. [doi: 10.1145/3233231]
46. Voigt P, von dem Bussche A. The EU General Data Protection Regulation (GDPR). 2017. [doi: 10.1007/978-3-319-57959-7]
47. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med* 2018;169:866–872. [doi: [10.7326/M18-1259
48. Scott IA, Van Der Vegt A, Lane P, McPhail S, Magrabi F. Achieving large-scale clinician adoption of AI-enabled decision support. *BMJ Health Care Inform* 2024;31:e100971. [doi: 10.1136/bmjhci-2023-100971]
49. Center for Democracy and Technology. Applying sociotechnical approaches to AI governance in practice. URL: https://cdt.org/insights/applying-sociotechnical-approaches-to-ai-governance-in-practice [accessed 2025-10-07]
50. MLOps Principles. URL: https://ml-ops.org/content/mlops-principles [accessed 2025-10-07]
51. MLflow Project. URL: https://mlflow.org/docs/latest [accessed 2025-10-07]
52. Debray TPA, et al. Transparent reporting of multivariable prediction models developed or validated using clustered data (TRIPOD-Cluster): explanation and elaboration. *BMJ* 2023;380:e071018. [doi: 10.1136/bmj-2022-071018]
53. OpenAIRE. ARGOS: create, link and share data management plans. URL: https://www.openaire.eu/argos [accessed 2025-10-07]
54. European Commission. Regulatory framework for artificial intelligence (AI Act). URL: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai [accessed 2025-10-07]

Abbreviations

AI: Artificial Intelligence
ALTAI: Assessment List for Trustworthy Artificial Intelligence
ARGOS: Aggregator for Open Science (European Open Science Cloud tool for data governance)
FDA: Food and Drug Administration
GDPR: General Data Protection Regulation
HLEG-AI: High-Level Expert Group on Artificial Intelligence
IP: Intellectual Property
ISO/IEC TS: International Organization for Standardization / International Electrotechnical Commission Technical Specification
LIME: Local Interpretable Model-Agnostic Explanations
MLflow: Machine Learning Flow (open-source platform for managing ML lifecycle)
MLOps: Machine Learning Operations
NIST: National Institute of Standards and Technology

OECD: Organisation for Economic Co-operation and Development
PD: Parkinson's Disease
SHAP: SHapley Additive exPlanations
TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
UX/UI: User Experience / User Interface
xAI: Explainable Artificial Intelligence

Supplementary Files

Multimedia Appendixes

Average importance score by ALTAI subgroup and full set of ALTAI checklist responses alongside the corresponding system-generated recommendations.
URL: http://asset.jmir.pub/assets/df391c298e21f6f9cb1b4e1377ee32ae.docx