Trustworthy AI for Safety- Critical Perception and Decision Systems

Shruti Kshirsagar
School of Computing, Wichita State University, Wichita, KS, USA
shruti.kshirsagar@wichita.edu

***Abstract*—The rapid advancement of artificial intelligence (AI) in safety-critical fields such as medical diagnostics, autonomous vehicles, disaster response, aerospace and infrastructure monitoring has made the challenge of trustworthiness not only an academic one, but an important real-world one. Systems that perceive, reason, and decide in high-stakes environments must satisfy a demanding set of properties: they must be explainable, robust under distribution shift and sensor noise, anomaly aware, human-centred, and verifiably reliable. This article provides the overview of trustworthy AI for safety-critical perception and decision systems, tracing the field from its motivating challenges through emerging technical foundations to open research questions and exemplary applications.We conclude that trustworthiness is a systems property, not a single algorithmic feature, and assert that realising it requires coordinated advances in explainability, uncertainty quantification, human–AI interaction design, failure detection, and domain-specific data governance. In addition, we outline a research agenda for the community and identify the cross-disciplinary bridges, spanning machine learning, human factors, and ethics, that must be built if AI is to become a genuinely dependable partner in life-critical domains**

***Index Terms*—Trustworthy AI, explainability, safety-critical systems, human-in-the-loop, anomaly detection, robust perception, deepfake detection, medical AI, disaster assessment.**

I. Introduction

In March 2019, the crash of Ethiopian Airlines Flight 302 highlighted a challenge that extends far beyond aviation. The tragedy was not simply a failure of software or hardware; it revealed a deeper systems problem involving automation, human judgment, and trust. Similar concerns now emerge across medicine, autonomous transportation, disaster management, cybersecurity, and critical infrastructure. As artificial intelligence assumes increasingly influential roles in decision-making, society faces a fundamental question: *"when should humans trust AI, and when should they not?"*

The question has become urgent because AI systems are no longer confined to recommendation engines or consumer applications. They now assist radiologists in detecting cancer [1], monitor patients in intensive care units [2], assess structural damage after natural disasters [3]–[6], support air traffic management, and guide autonomous vehicles through dynamic environments [7]. In these contexts, errors can result not merely in inconvenience but in injury, loss of life, environmental harm, or large-scale economic disruption.

For decades, the dominant objective in artificial intelligence research was improving predictive accuracy. Benchmark performance became the primary measure of progress. Yet experience has repeatedly demonstrated that high accuracy alone does not guarantee safe deployment. Models that achieve impressive results in laboratory settings often perform unpredictably when tested with noisy sensors, changing environments, unforeseen operating conditions, or novel situations that differ from their training data [5], [8], [9].

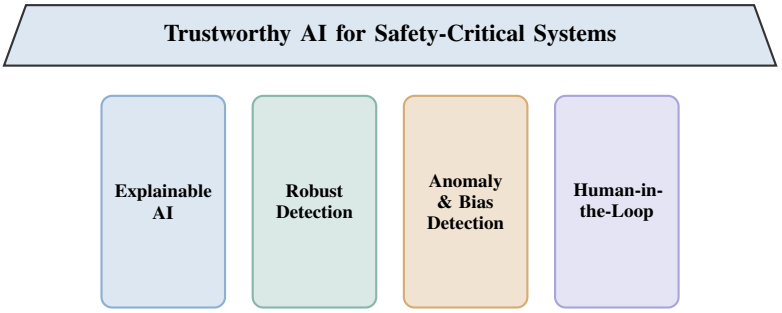

Fig. 1. The four mutually reinforcing pillars of trustworthy AI for safety-critical systems.

The challenge is especially evident in safety-critical domains. An autonomous vehicle may correctly identify pedestrians under normal conditions yet fail during heavy rain or unusual lighting [7]. A clinical decision-support system may achieve expert-level diagnostic performance on curated datasets while producing unreliable recommendations when deployed across different hospitals [10]. Disaster assessment algorithms may struggle when incorporated with disaster types or geographic regions not represented in training data a challenge directly addressed in recent work on building damage detection [4], [6]. In each case, the problem is not merely one of accuracy but of trustworthiness.

Trust in an AI system is multidimensional [11]. It encompasses the ability of a human operator to understand why a decision was made (explainability), confidence that performance will not degrade catastrophically under novel or corrupted inputs (robustness), timely awareness of system failures or anomalies (fault detection), and a design philosophy that keeps human judgment authoritative when uncertainty is high (human-in-the-loop) [12]. As Fig. 1 illustrates, these pillars are mutually reinforcing: explainability supports better human oversight, which in turn provides correction signals that improve robustness over time.

The scale of this concern is quantifiable. A 2023 OECD survey revealed that AI-related incidents in high-risk industries have surged exponentially, by more than 200% since 2020, with healthcare and transportation accounting for the largest

share [13]. Models that attain 95% accuracy on curated test sets often drop below 70% when faced with the noise, occlusion, and distributional shifts that characterise authentic operational environments [8]. The performance drop is not simply an engineering inconvenience: in a patient monitoring system a 25-point accuracy decrease directly results in missed diagnoses; in an autonomous vehicle it leads to collisions; in a disaster response system it causes misallocated resources and preventable casualties.

The urgency is also reflected in the policy landscape. The EU AI Act (2024) classifies medical devices, critical infrastructure, and autonomous vehicles as high-risk AI systems subject to mandatory conformity assessment [14]. The US NIST AI Risk Management Framework [13] and the IEEE 7000 series on ethically aligned design similarly call for systematic trustworthiness evaluation. Yet the research community lacks consensus on operationalising these requirements into deployable engineering practice.

This article aims to bridge that gap. Section II examines the key challenges associated with deploying AI in safety-critical environments. Section III reviews the technical foundations that underpin trustworthy AI, including explainability, robustness, uncertainty quantification, anomaly detection, and human-AI collaboration. Section IV discusses the ongoing shift from isolated AI components toward integrated trustworthy AI systems. Section V highlights representative applications in healthcare, assistive technologies and human-computer interaction, and autonomous and disaster-response systems. Section VI outlines the major open challenges and future research directions shaping the next generation of trustworthy AI. Finally, Section VII concludes the article.

II. Core Challenges in Safety-Critical AI

Deploying AI in safety-critical contexts exposes a cluster of challenges that are qualitatively different from those encountered in consumer applications.

A. The Opacity Problem

Modern deep neural networks achieve remarkable accuracy by learning highly nonlinear, high-dimensional representations that resist human interpretation [15]. In a consumer context, opacity is disruptive; in a clinical or aviation context, it can be fatal. A radiologist cannot act on a lung-cancer prediction they cannot interrogate. A pilot cannot trust an autopilot whose reasoning is invisible. Explainability is therefore not a ”nice to have” feature but a prerequisite for deployment [16].

B. Distribution Shift and Sensor Degradation

Real-world sensors are noisy, occluded, and subject to calibration drift. Models trained on clean, curated datasets often fail when deployed in the field [8]. Across five representative safety-critical benchmarks, mean accuracy under common corruptions (noise, blur, weather, digital artefacts) drops by 18-41 percentage points relative to clean-data performance. This has been observed in medical imaging workflows [10], autonomous driving stacks [7], and disaster remote sensing pipelines [4]. Robust perception under degraded inputs is therefore a first-order research priority.

C. Failure Detection and Silent Errors

A distinguishing feature of safety critical AI is that silent failures confidently wrong predictions are more dangerous than abstentions. Systems must know what they do not know [17]. Anomaly detection and out-of-distribution (OOD) recognition are therefore integral components, not post-hoc additions.

D. Human–AI Trust Calibration

Evidence from aviation [18], medicine [19], and nuclear operations consistently shows that humans either over-trust automated systems (automation bias) or under-trust them after a single failure. Neither extreme is safe. Calibrated trust proportional to actual system reliability in context requires explicit interaction design, uncertainty communication, and feedback mechanisms [20]. Human-in-the-loop frameworks enable experts to review, validate, and override AI recommendations when necessary. Establishing clear accountability between human operators and automated systems is essential for trustworthy deployment.

E. Data Scarcity and Label Noise in Critical Domains

Safety-critical events are, by definition, rare. Labelled failure cases, disease positives, and disaster imagery are scarce relative to the distributional diversity of deployment conditions. Label noise is compounded by inter-annotator disagreement among domain experts [21]. These data challenges demand specialised learning strategies beyond standard supervised learning.

F. Bias in Data and Models

Bias in AI systems refers to systematic errors introduced by unrepresentative data, flawed assumptions, or learned patterns that lead to unfair, inaccurate, or inconsistent outcomes across different populations, environments, or operating conditions. AI systems learn from historical data, which may contain demographic, geographic, or institutional biases. Such biases can result in unfair or inaccurate predictions for underrepresented populations or operating conditions [22]. Ensuring representative datasets and fairness-aware model development remains a critical challenge, particularly in healthcare and disaster response where bias directly translates to differential harm.

III. Technical Foundations for Trustworthy AI

The research community has responded to the challenges in Section II with a growing toolkit. This section surveys the principal technical foundations that underpin our workshop themes.

A. Explainability Methods

Post-hoc explainability methods decompose a trained model’s predictions into human-interpretable components. LIME [23] approximates the local decision boundary with an interpretable surrogate. SHAP [24] uses Shapley values to assign globally consistent feature attributions. Grad-CAM [25] localises salient image regions for convolutional networks, with direct

clinical utility in radiology and pathology. Concept-based explanations [26] align internal representations with human-defined concepts, enabling domain experts to audit model reasoning using their own vocabulary. Intrinsically interpretable architectures decision trees, rule lists, case-based reasoners, and monotone neural networks [16] avoid the fidelity gap of post-hoc methods but at some expressivity cost. Recent work on self-explaining neural networks and prototype networks [27] attempts to retain deep-learning accuracy while building explanations into the architecture itself.

A critical but underappreciated axis is explanation fidelity versus explanation plausibility [15]. Plausible explanations align with domain expert intuition; faithful explanations accurately reflect the model's actual computation. These properties are not equivalent a post-hoc method can produce plausible but unfaithful explanations, which may reassure a clinician while concealing a spurious reasoning shortcut. Evaluation frameworks that separately audit fidelity and plausibility, using held-out causal interventions or adversarial explanation stress-tests [28], are therefore essential components of a trustworthiness assessment pipeline.

B. Uncertainty Quantification

Reliable uncertainty estimates are a prerequisite for safe deployment. Bayesian deep learning [29] provides principled posterior estimates via Monte Carlo dropout at test time. Deep ensembles [30] produce well-calibrated predictive distributions at modest computational overhead. Conformal prediction [31] provides distribution-free, finite-sample coverage guarantees an increasingly attractive property for regulatory submissions. Post-hoc calibration via temperature scaling [32] corrects overconfidence in pre-trained models without retraining.

C. Robust Perception under Noise and Incomplete Data

Adversarial training [33] augments datasets with worst-case perturbations to improve certified robustness. Data augmentation strategies including MixUp, CutMix, and AutoAugment generalise training distributions. Self-supervised pre-training on large unlabelled corpora builds robust representations transferable to low-resource safety-critical tasks [34]. Sensor fusion architectures that aggregate redundant modalities (RGB + depth + LiDAR + radar) provide graceful degradation when individual sensors fail [7].

D. Anomaly Detection and OOD Awareness

Classical statistical process control provides interpretable anomaly scores in low-dimensional settings. Deep one-class classifiers [35] and energy-based models [36] scale to complex sensory data. Foundation-model embeddings combined with lightweight OOD heads [37] offer an efficient route to anomaly detection when labelled anomalies are unavailable. For physiological monitoring, transformer-based sequence models detect subtle precursor patterns preceding cardiac or neurological events [2].

E. Human-in-the-Loop Learning

Active learning selectively queries human experts for labels on the most informative samples, dramatically reducing annotation cost in label-scarce domains [38]. Reinforcement learning from human feedback (RLHF) [39] aligns model behaviour with expert preferences, while interactive machine learning frameworks [40] allow domain experts to correct model errors in real time. Mixed-initiative systems explicitly partition decision authority between human and AI based on estimated confidence and task context [41].

A key empirical finding is that the benefit of human oversight is highly task- and confidence-dependent [18]. When AI confidence is high and task complexity is low, human intervention adds latency without improving accuracy. Conversely, at the tails of the confidence distribution precisely where errors are most consequential human judgment consistently outperforms autonomous AI. The design implication is a dynamic handoff boundary: the system should automatically escalate to human review when uncertainty exceeds a domain-calibrated threshold, and progressively automate as accumulated evidence builds confidence in a particular decision context.

IV. New Paradigm: The Trustworthy AI Systems Era

The convergence of the techniques described in Section III is not merely an incremental advance it constitutes a paradigm shift in how AI for safety-critical systems is conceived, designed, and evaluated.

A. Trustworthiness as a Measurable Property

A prerequisite for the paradigm shift is agreeing on what to measure. The community has converged on several key metrics: Expected Calibration Error (ECE) and AUROC for uncertainty and OOD detection; mean Corruption Error (mCE) and certified robustness radius for perception; faithfulness and plausibility scores for explainability; and Safety Case Coverage for verified deployment [8], [42], [43]. The absence of a unified benchmark integrating all five dimensions remains a key open challenge identified in Section VI.

B. From Components to Systems

The prevailing research paradigm treats explainability, robustness, and human factors as separate workstreams. The emerging paradigm recognises them as co-dependent properties of an integrated system. Fig. 2 illustrates this ecosystem view: perception, reasoning, explanation, monitoring, and human interaction form a closed-loop system with continuous feedback rather than a one-shot inference pipeline.

C. Foundation Models as Trustworthy Backbones

Large pre-trained models vision transformers [44], CLIP [45], and domain-adapted models such as Med-PALM 2 [46] and GeoSAM [47] provide robust, transferable representations that reduce the data scarcity challenge. However, their scale introduces new interpretability and auditability challenges. The emerging response is retrieval-augmented and evidence-grounded prediction [48], where predictions are linked to specific supporting evidence that a domain expert can verify.

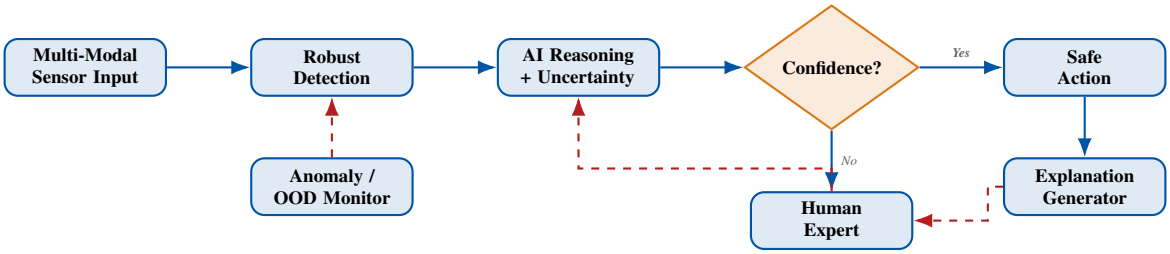

Fig. 2. Closed-loop trustworthy AI ecosystem. Perception, reasoning, explanation, anomaly monitoring, and human oversight interact continuously rather than as sequential pipeline stages.

D. LLM-Driven Decision Support

Large language models (LLMs) are beginning to serve as cognitive mediators in safety-critical workflows [49]: translating clinical notes into structured risk scores, summarising multi-sensor alarm streams for operators, or generating natural-language justifications for autonomous system decisions. Multi-agent architectures [50] allow specialised AI agents to collaborate on complex diagnostic or planning tasks, with explicit debate and verification steps that expose disagreement before it reaches the human operator a direct analogue of crew resource management in aviation.

E. Regulatory Alignment and Standards

The paradigm shift extends beyond technology. Regulatory bodies now require AI developers to produce technical documentation demonstrating conformity with safety requirements throughout the AI lifecycle. Concepts such as AI safety cases structured arguments that a system is acceptably safe for a defined operating context [43] are migrating from the nuclear and aerospace industries into medical devices and autonomous vehicles. Standards bodies (ISO/IEC 42001, IEC 62304, DO-178C) are actively developing AI-specific annexes. The community must therefore produce not only accurate models but arguable ones.

V. Applications of Trustworthy AI in Safety-Critical Domains

Trustworthy AI is increasingly being deployed across a wide range of safety-critical domains where reliability, transparency, and human oversight are essential. While application requirements differ across sectors, common trustworthiness challenges include explainability, uncertainty quantification, robustness to distribution shift, anomaly awareness, and effective human-AI collaboration. We highlight three major application areas where these principles are particularly important.

A. Trustworthy AI in Healthcare and Medical Decision Support

Healthcare is among the most demanding application domains for trustworthy AI because prediction errors can directly impact patient safety and clinical outcomes. Recent advances in deep learning have demonstrated expert-level performance in radiology, pathology, and physiological signal analysis [1], [2]. However, successful clinical deployment requires more than predictive accuracy. Models must provide interpretable explanations, calibrated confidence estimates, and robustness across institutions, patient populations, and imaging devices [10], [17].

Explainable AI techniques such as SHAP, Grad-CAM, and concept-based explanations enable clinicians to understand model predictions and identify potential failure modes [25], [26]. Likewise, uncertainty-aware inference allows systems to flag ambiguous cases for expert review rather than producing overconfident predictions. Federated learning has emerged as a promising framework for addressing data-sharing and privacy challenges while enabling collaborative model development across healthcare institutions [52].

Examples of trustworthy healthcare AI include multimodal physiological monitoring systems integrating electrocardiogram (ECG) and phonocardiogram (PCG) signals for cardiac assessment [53], transformer-based analysis of neurological and physiological signals [2], and AI-assisted brain tumor infiltration prediction systems that combine predictive performance with explainable decision support [54]. In all cases, human clinicians remain central to the decision-making process, highlighting the importance of trust calibration and human oversight.

B. Trustworthy AI for Assistive Technologies and Human-Computer Interaction

Beyond automation, AI increasingly functions as an assistive technology that augments human decision making in healthcare, cybersecurity, industrial operations, and education. In these settings, AI serves as a cognitive partner rather than an autonomous decision maker, requiring effective interaction mechanisms that support transparency, collaboration, and user trust [12], [20].

Human-centered AI systems incorporate explainability, uncertainty communication, and interactive feedback loops that allow users to understand, validate, and override AI recommendations when necessary [19], [55]. Mixed-initiative systems dynamically allocate authority between human operators and AI based on task complexity, operational context, and confidence estimates [41]. Recent advances in large language models (LLMs) further expand these capabilities by enabling natural-language explanations, decision summaries, and conversational decision support interfaces [49].

Media authenticity and deepfake detection represent another emerging area of trustworthy assistive AI. As synthetic media become increasingly realistic, AI-based forensic tools help users assess the credibility of visual and audio content [56],

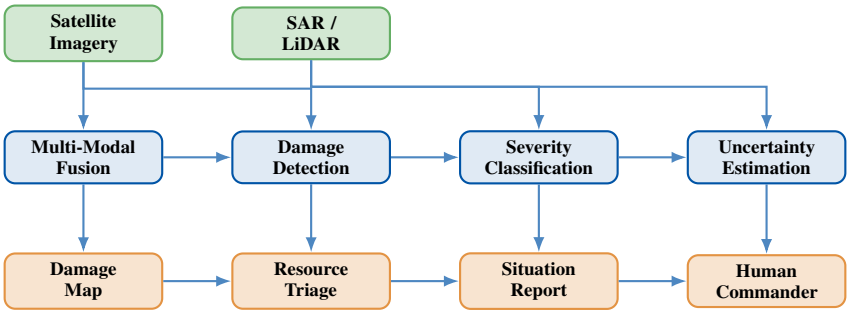

Fig. 3. Disaster AI pipeline: satellite imagery and SAR/LiDAR data are fused, processed for damage detection and severity classification with uncertainty estimation, and surfaced to human commanders as actionable situation reports [3], [51].

[57]. Recent work has explored phonetic and self-supervised speech representations for detecting synthetic audio while simultaneously evaluating fairness across demographic groups [58]. These systems exemplify how trustworthy AI can support informed human judgment rather than replace it.

C. Trustworthy AI for Disaster Response, Infrastructure Monitoring, and Autonomous Systems

Safety-critical operational environments require AI systems capable of perceiving, reasoning, and acting under uncertainty. Autonomous vehicles, disaster response platforms, and infrastructure monitoring systems must operate reliably despite sensor degradation, adverse weather conditions, incomplete information, and unforeseen operating scenarios [7], [8].

Recent advances in multimodal sensing, sensor fusion, and uncertainty-aware perception have improved the robustness of these systems. Autonomous vehicle platforms increasingly integrate cameras, LiDAR, radar, and depth sensors to achieve graceful degradation when individual sensing modalities fail [7]. At the same time, anomaly detection and out-of-distribution (OOD) monitoring techniques enable systems to recognize unfamiliar operating conditions and request human intervention when necessary [35], [36].

In disaster management, trustworthy AI supports rapid situational awareness by combining satellite imagery, social-media data, and meteorological information [51]. Building damage assessment systems developed using datasets such as xBD provide critical information for emergency response and resource allocation following natural disasters [3]. Recent research has focused on improving robustness across geographic regions and disaster types while incorporating uncertainty estimation into operational workflows [5], [6]. Similar approaches are increasingly being applied to infrastructure monitoring, where AI systems detect structural degradation, equipment failures, and other anomalies before they escalate into catastrophic events.

Across these domains, trustworthy AI depends not only on accurate prediction but also on the ability to communicate uncertainty, detect failures, and maintain meaningful human oversight throughout the decision-making process.

VI. Open Challenges and Future Directions

Despite remarkable advances in artificial intelligence, achieving trustworthy AI for safety-critical systems remains an open research challenge. Across healthcare, assistive technologies, autonomous systems, disaster response, and critical infrastructure, successful deployment requires more than high predictive accuracy. Systems must provide reliable uncertainty estimates, remain robust under changing operating conditions, support meaningful human oversight, and satisfy increasingly stringent regulatory requirements. Addressing these challenges will require coordinated advances across machine learning, human factors, systems engineering, and public policy.

A. Robustness and Adaptation Under Distribution Shift

Real-world environments are dynamic and continuously evolving. Medical protocols change, sensors degrade, infrastructure ages, and new threat patterns emerge. Models trained on historical data often experience significant performance degradation when deployed under conditions that differ from their training distributions [8], [17]. Future trustworthy AI systems must be capable of adapting to changing environments while preserving previously validated behaviors. Research in continual learning, domain adaptation, test-time adaptation, and robust representation learning offers promising directions, but practical deployment remains limited by challenges such as catastrophic forgetting and the absence of reliable re-certification mechanisms [59].

B. Causal and Explainable Reasoning

Many current AI systems rely primarily on statistical correlations rather than causal understanding. While such models may achieve high predictive accuracy, they often struggle to generalize under distributional shifts and provide limited support for human decision-making. Future research must move beyond post-hoc explanations toward models capable of reasoning about cause-and-effect relationships and generating actionable counterfactual explanations [60]. Integrating causal reasoning with foundation models may enable systems that not only predict outcomes but also explain why decisions are made and how alternative actions could influence future outcomes.

C. Human–AI Collaboration and Trust Calibration

The future of trustworthy AI is unlikely to be fully autonomous. Instead, AI systems will increasingly operate as collaborative partners that support human expertise in

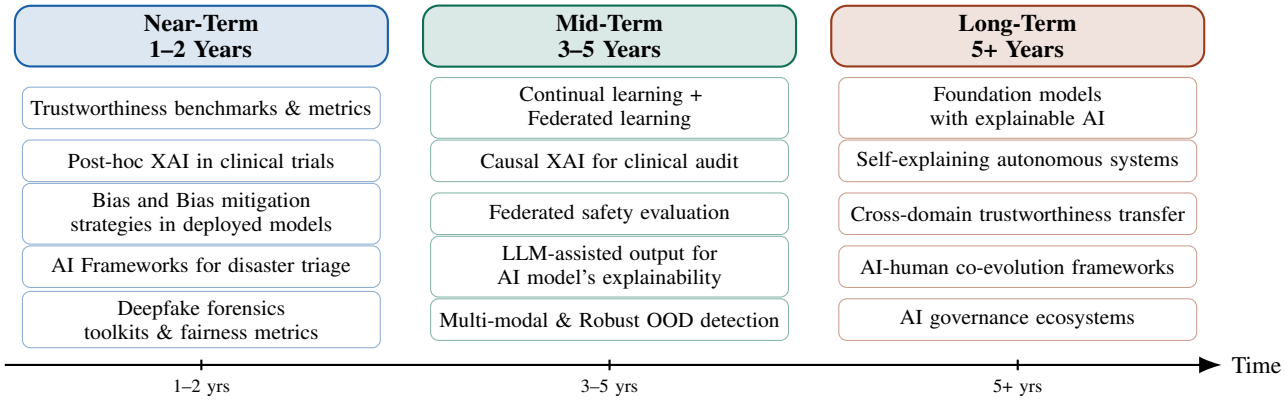

Fig. 4. Three-horizon research roadmap for trustworthy AI in safety-critical systems, spanning near-term deployable solutions through long-term foundational research goals.

complex decision-making environments. A central challenge is achieving appropriate trust calibration, ensuring that users neither over-rely on nor underutilize AI recommendations [18], [20]. Future research should focus on adaptive interfaces, uncertainty communication, mixed-initiative decision-making, and personalized explanation strategies that account for user expertise, workload, and situational context. Understanding how humans interact with increasingly capable AI systems remains one of the most important interdisciplinary challenges in the field.

D. Evaluation Frameworks and Benchmarking

Progress in trustworthy AI is hindered by the absence of standardized evaluation methodologies. Existing metrics for explainability, robustness, calibration, fairness, and anomaly detection are often evaluated independently, making it difficult to assess overall system trustworthiness [28]. Future efforts should focus on developing comprehensive benchmark suites that evaluate multiple dimensions of trustworthiness under realistic deployment conditions. Similar to crash-testing standards in the automotive industry, trustworthy AI will require rigorous stress-testing protocols capable of assessing performance under uncertainty, adversarial conditions, and rare edge cases.

E. Governance, Ethics, and Regulatory Alignment

As AI systems become increasingly integrated into critical societal functions, governance and regulatory considerations will play a growing role in deployment decisions. Emerging frameworks such as the NIST AI Risk Management Framework, the EU AI Act, and AI-specific safety standards emphasize accountability, transparency, risk management, and human oversight [13], [14]. Future research must therefore address not only technical performance but also compliance, auditability, fairness, privacy, and accountability throughout the AI lifecycle. Building trustworthy AI will require close collaboration among researchers, industry practitioners, policymakers, and domain experts.

F. Research Roadmap

Figure 4 summarizes a three-horizon roadmap for advancing trustworthy AI in safety-critical systems. Near-term efforts should focus on establishing standardized trustworthiness metrics, improving explainability and uncertainty estimation, and developing practical deployment frameworks for high-risk applications. Mid-term priorities include continual learning, causal reasoning, federated evaluation, and robust multimodal AI systems capable of operating under distributional shift. Looking further ahead, long-term research aims to enable self-explaining, verifiable, and adaptive AI systems that collaborate effectively with human stakeholders while maintaining transparency, accountability, and safety. Achieving this vision will require a shift from optimizing isolated models toward engineering trustworthy AI ecosystems that integrate technical, human, and organizational considerations.

VII. Conclusion

Trustworthy AI for safety-critical systems represents one of the most important challenges at the intersection of artificial intelligence, human-centered design, and systems engineering. As AI technologies become increasingly integrated into healthcare, assistive technologies, autonomous systems, disaster response, and critical infrastructure, ensuring that these systems are explainable, robust, reliable, and aligned with human values is essential. This article has highlighted the technical foundations, application domains, and emerging challenges that define the field, emphasizing that trustworthiness is not a single algorithmic property but a systems-level characteristic that depends on effective uncertainty management, human oversight, continual adaptation, and responsible governance. Moving forward, advances in verification, causal reasoning, human-AI collaboration, and standardized evaluation frameworks will be critical to building AI systems that not only perform accurately but also operate safely, transparently, and accountably in real-world environments. Ultimately, the goal of trustworthy AI is not to replace human decision makers but to create dependable and collaborative intelligent systems that enhance human capabilities while preserving safety, trust, and societal well-being.

Shruti Kshirsagar is an Assistant Professor and Graduate Coordinator for the MS in Data Science program in the School of Computing at Wichita State University (WSU), Wichita, KS, USA. She received her Ph.D. in Telecommunications from the Institut National de la Recherche Scientifique (INRS), Montréal, Canada. She leads the SoundMind Neurovision Innovation Lab at WSU, where her research spans trustworthy and explainable AI, medical image processing, sleep analysis, deepfake detection for audio and multimodal signals, affective computing, and AI for disaster response. Her work is supported by NSF grants. She serves as Workshop Organiser for the IEEE SMC 2026 Workshop on Trustworthy AI for Safety-Critical Systems, Associate Editor for IEEE EMBC, IEEE ICHMS, and IEEE SMC, reviewer for the National Science Foundation. Contact: *shruti.kshirsagar@wichita.edu*

References

[1] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine," *Nature Medicine*, vol. 28, no. 1, pp. 31–38, 2022.

[2] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1589–1604, 2018.

[3] R. Gupta, B. Goodman, N. Patel, R. Hosfelt, S. Sajeev, E. Heim, J. Doshi, K. Lucas, H. Y. Kim, and M. Shipman, "Creating xBD: A dataset for assessing building damage from satellite imagery," in *Proceedings of the IEEE/CVF CVPR Workshops*, 2019. [Online]. Available: https://xview2.org/paper

[4] B. C. Reddy, S. Kshirsagar, R. Bagai, and A. Dutta, "Towards robust building damage detection: Leveraging augmentation and domain adaptation," in *Proceedings of the IEEE GreenTech Conference*, 2025, nSF EPSCoR funded.

[5] S. Kshirsagar, B. Chandra, U. Tallal, R. Bagai, and A. Dutta, "Geographic bias analysis and cross-domain generalization in deep learning-based building damage assessment," *Remote Sensing*, vol. 18, no. 10, p. 1529, 2026.

[6] A. Mouradi and S. Kshirsagar, "Robust building damage detection in cross-disaster settings using domain adaptation," *arXiv e-prints*, pp. arXiv–2603, 2026.

[7] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuScenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 621–11 631.

[8] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *arXiv preprint arXiv:1903.12261*, 2019. [Online]. Available: https://arxiv.org/abs/1903.12261

[9] S. Kshirsagar and T. H. Falk, "Quality-aware bag of modulation spectrum features for robust speech emotion recognition," *IEEE Transactions on Affective Computing*, 2022.

[10] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine*, vol. 17, no. 1, pp. 1–9, 2019.

[11] A. Jacovi, A. Marasovic, T. Miller, and Y. Goldberg, "Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI," in *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021, pp. 624–635.

[12] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, and E. Horvitz, "Guidelines for human-AI interaction," in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–13.

[13] National Institute of Standards and Technology (NIST), "AI risk management framework (AI RMF 1.0)," U.S. Department of Commerce, Tech. Rep. NIST AI 100-1, 2023. [Online]. Available: https://doi.org/10.6028/NIST.AI.100-1

[14] European Parliament and Council, "Regulation (EU) 2024/1689 of the european parliament and of the council: Artificial intelligence act," Official Journal of the European Union, Tech. Rep., 2024. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

[15] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[16] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[17] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, B. Lakshminarayanan, J. Snoek, Z. Ghahramani, and J. V. Dillon, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019. [Online]. Available: https://papers.nips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html

[18] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Transactions on Systems, Man, and Cybernetics — Part A: Systems and Humans*, vol. 30, no. 3, pp. 286–297, 2000.

[19] C. J. Cai, E. Reif, N. Hegde, J. Hipp, B. Kim, D. Smilkov, M. Wattenberg, M. Lungren, M. H. Chou, M. Ghassemi, and M. Terry, "Human centered tools for coping with imperfect algorithms during medical decision-making," in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–14.

[20] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[21] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Medical Image Analysis*, vol. 65, p. 101759, 2020.

[22] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023. [Online]. Available: https://fairmlbook.org

[23] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[24] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 4765–4774. [Online]. Available: https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020.

[26] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, "Interpretability beyond classification: Quantitative testing with concept activation vectors (TCAV)," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2018, pp. 2668–2677. [Online]. Available: https://proceedings.mlr.press/v80/kim18d.html

[27] C. Chen, O. Li, D. Tao, A. J. Barnett, C. Rudin, and J. K. Su, "This looks like that: Deep learning for interpretable image recognition," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019. [Online]. Available: https://papers.nips.cc/paper/2019/hash/adf7ee2dcf142b0e11888e72b43fcb75-Abstract.html

[28] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond accuracy: Behavioral testing of NLP models with CheckList," in *Proceedings of the ACL*, 2020, pp. 4902–4912.

[29] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2016, pp. 1050–1059. [Online]. Available: https://proceedings.mlr.press/v48/gal16.html

[30] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017. [Online]. Available: https://papers.nips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html

[31] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," *arXiv preprint arXiv:2107.07511*, 2021. [Online]. Available: https://arxiv.org/abs/2107.07511

[32] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2017, pp. 1321–1330. [Online]. Available: https://proceedings.mlr.press/v70/guo17a.html

[33] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: https://openreview.net/forum?id=rJzIBfZAb

[34] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2020,

pp. 1597–1607. [Online]. Available: https://proceedings.mlr.press/v119/chen20j.html

[35] L. Ruff, R. A. Vandermeulen, N. Görnitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2018, pp. 4393–4402. [Online]. Available: https://proceedings.mlr.press/v80/ruff18a.html

[36] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 21 464–21 475. [Online]. Available: https://papers.nips.cc/paper/2020/hash/f5496252609c43eb8a3d147ab9b9c006-Abstract.html

[37] Y. Ming, Y. Sun, O. Dia, and Y. Li, "Delving into out-of-distribution detection with vision-language representations," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022.

[38] B. Settles, "Active learning literature survey," *University of Wisconsin–Madison, Computer Sciences Technical Report 1648*, 2009. [Online]. Available: http://burrsettles.com/pub/settles.activelearning.pdf

[39] P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017. [Online]. Available: https://papers.nips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html

[40] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning," vol. 35, no. 4, 2014, pp. 105–120.

[41] E. Horvitz, "Principles of mixed-initiative user interfaces," in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 1999, pp. 159–166.

[42] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," in *arXiv preprint arXiv:1702.08608*, 2017. [Online]. Available: https://arxiv.org/abs/1702.08608

[43] R. Hawkins, C. Paterson, C. Picardi, Y. Jia, R. Calinescu, and I. Habli, "Guidance on the assurance of machine learning in autonomous systems (AMLAS)," in *arXiv preprint arXiv:2102.01564*, 2021. [Online]. Available: https://arxiv.org/abs/2102.01564

[44] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[45] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[46] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[47] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4015–4026.

[48] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 9459–9474. [Online]. Available: https://papers.nips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[49] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar, "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, no. 7956, pp. 259–265, 2023.

[50] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*, 2023, pp. 1–22.

[51] F. Alam, H. Sajjad, M. Imran, and F. Ofli, "MEDIC: A multi-task learning dataset for disaster image classification," in *arXiv preprint arXiv:2108.12828*, 2021. [Online]. Available: https://arxiv.org/abs/2108.12828

[52] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein *et al.*, "The future of digital health with federated learning," *npj Digital Medicine*, vol. 3, no. 1, pp. 1–7, 2020.

[53] M. Thu and S. Kshirsagar, "A machine learning approach for integrating phonocardiogram and electro-cardiogram data for heart sound detection," in *Proceedings of the International Conference on Signal Processing (ICSP)*, Germany, 2023.

[54] S. M. A. Hossain and S. Kshirsagar, "InfiltrNet: Dual-branch CNN-Transformer architecture for brain tumor infiltration risk prediction," *arXiv preprint arXiv:2605.02230*, 2025. [Online]. Available: https://arxiv.org/abs/2605.02230

[55] B. Shneiderman, *Human-Centered AI*. Oxford University Press, 2022.

[56] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, 2020.

[57] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1–11.

[58] A. Fursule, S. Kshirsagar, and A. R. Avila, "Gender fairness in audio deepfake detection: Performance and disparity analysis," in *Proceedings of the IEEE Conference on Artificial Intelligence (CAI)*, 2026.

[59] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–71, 2019.

[60] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021.