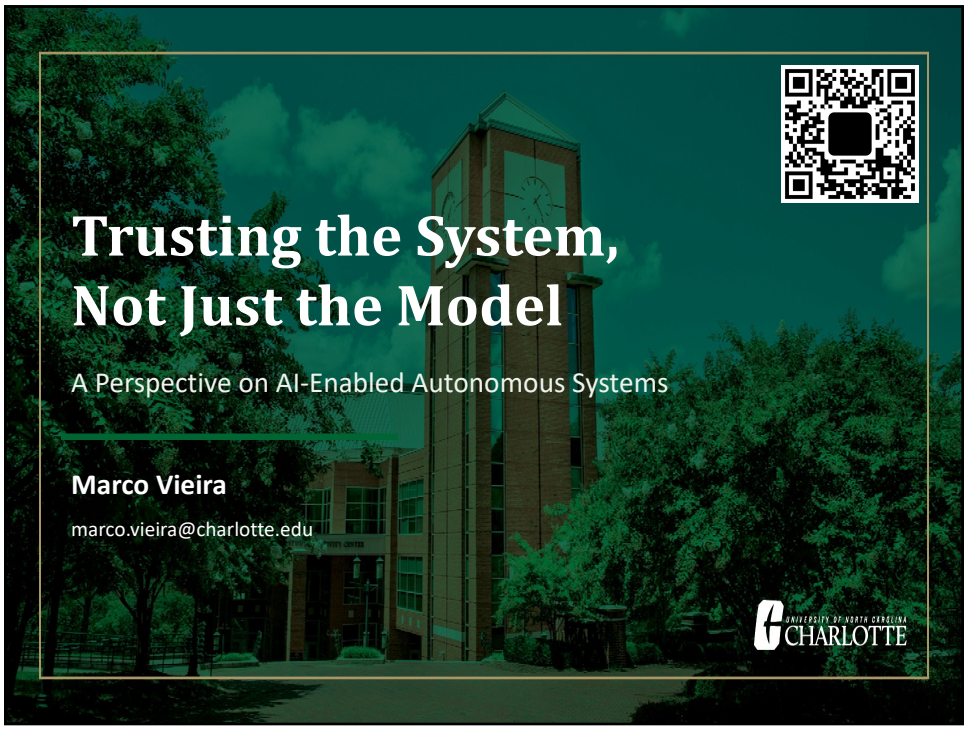
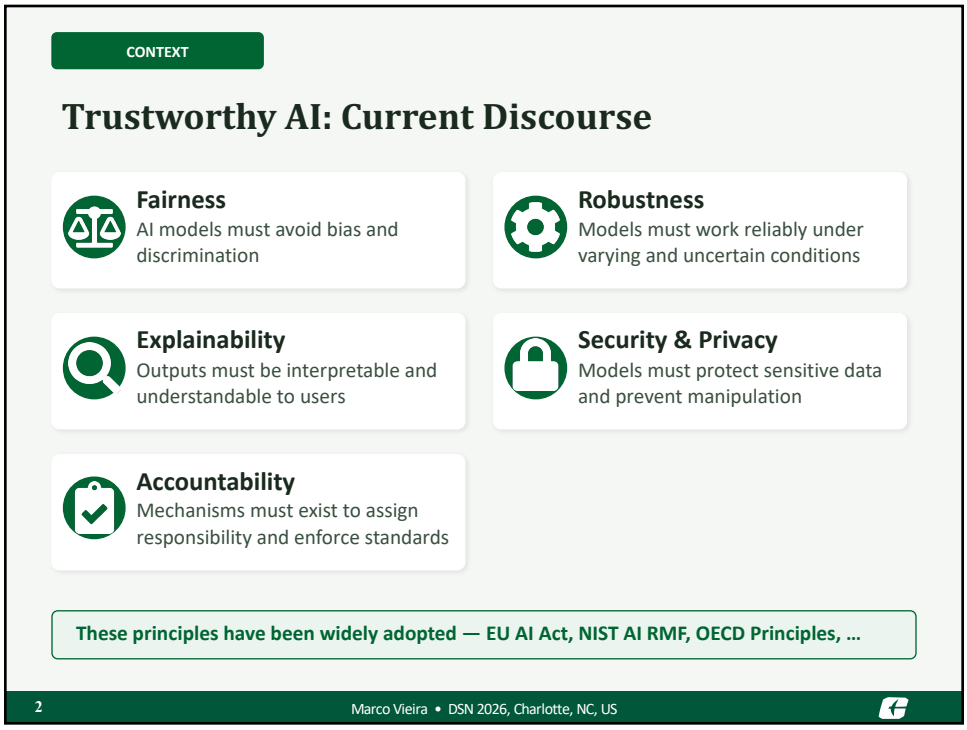

Trusting the System,
Not Just the Model
A Perspective on AI-Enabled Autonomous Systems
Marco Vieira
marco.vieira@charlotte.edu
UNIVERSITY OF NORTH CAROLINA
CHARLOTTE

1


CONTEXT
Trustworthy AI: Current Discourse
Fairness
AI models must avoid bias and discrimination
Robustness
Models must work reliably under varying and uncertain conditions
Explainability
Outputs must be interpretable and understandable to users
Security & Privacy
Models must protect sensitive data and prevent manipulation
Accountability
Mechanisms must exist to assign responsibility and enforce standards
These principles have been widely adopted — EU AI Act, NIST AI RMF, OECD Principles, …
2
Marco Vieira • DSN 2026, Charlotte, NC, US

2

THE PROBLEM

Fundamental Blind Spot

Why the focus on algorithms?

- **Opacity**: internal workings are difficult to inspect or explain
- **Bias**: emerging from training data in subtle, hard-to-detect ways
- **Verification**: formal guarantees much harder than for traditional software

A well-designed AI model may still produce harmful outcomes:

- Flawed data pipeline
- Insecure storage system
- Decision-making lacking human oversight

Such a narrow focus ignores a key aspect:

AI does not operate in isolation!

3

DEFINITION

What Trustworthiness Actually Means

The ability of the system to demonstrate that it fulfills a set of attributes within its intended goal

Lifecycle

Extend throughout the entire AI lifecycle: from design and training to deployment, monitoring, and decommissioning

Dynamic

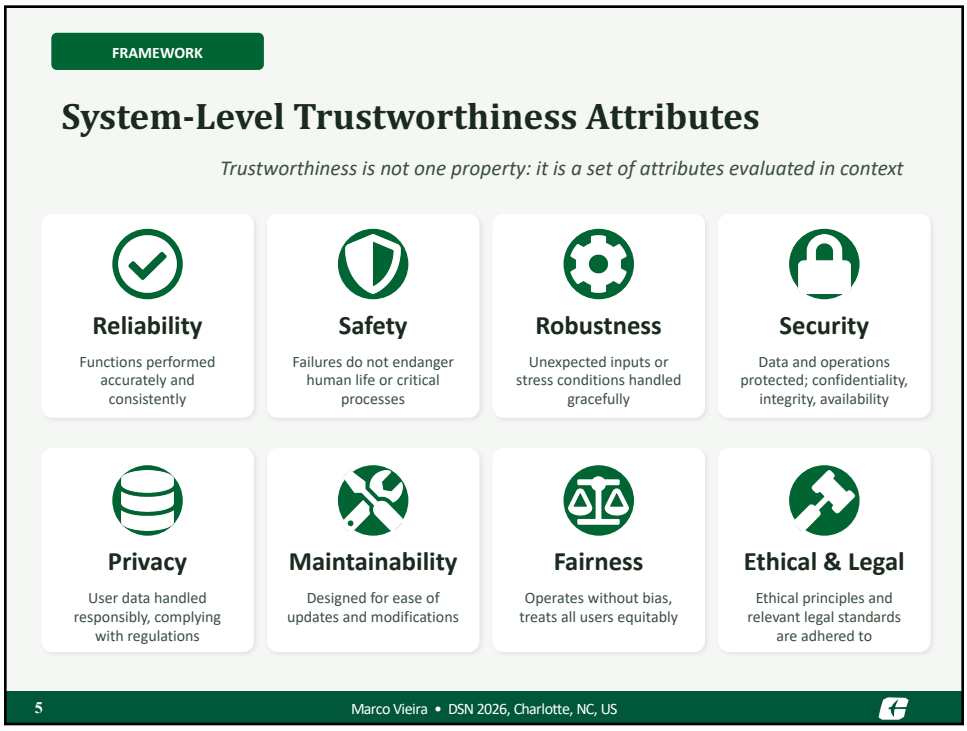
A model may be fair at training but become unreliable on outdated data or in a poorly monitored environment

Context-Specific

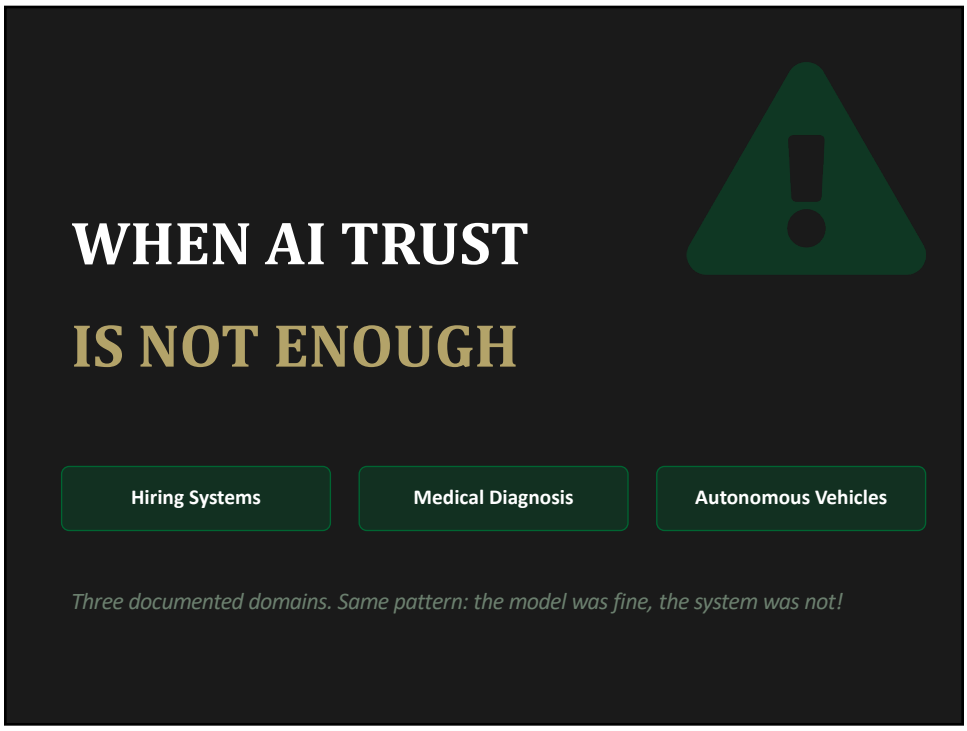
The relevant attributes vary across applications, domains, and stakeholder expectations

Without system-level understanding → systemic failures • governance gaps • human errors

4


FRAMEWORK
System-Level Trustworthiness Attributes
Trustworthiness is not one property: it is a set of attributes evaluated in context
Reliability
Functions performed accurately and consistently
Safety
Failures do not endanger human life or critical processes
Robustness
Unexpected inputs or stress conditions handled gracefully
Security
Data and operations protected; confidentiality, integrity, availability
Privacy
User data handled responsibly, complying with regulations
Maintainability
Designed for ease of updates and modifications
Fairness
Operates without bias, treats all users equitably
Ethical & Legal
Ethical principles and relevant legal standards are adhered to
5
Marco Vieira • DSN 2026, Charlotte, NC, US

5


WHEN AI TRUST
IS NOT ENOUGH
Hiring Systems
Medical Diagnosis
Autonomous Vehicles
Three documented domains. Same pattern: the model was fine, the system was not!

6

CASE 1 OF 3

AI Hiring Systems

Amazon Recruiting Tool

- The model excluded gendered terms, but penalized women
- Biased historical training data overrode model-level fairness
- The model reflected systemic inequality even though gender was not an explicit input

LinkedIn Recruiter

- Designed to rank candidates fairly, the system adapted to biased recruiter behavior
- Feedback loops twisted fairness over time without any direct gender signal

Facebook Ads Delivery

- Despite neutral targeting inputs, the delivery system showed janitorial roles more often to women and engineering roles to men
- Engagement-based optimization created discrimination through system dynamics

Pattern: Bias re-entered through data pipelines & human override.
The failure was not in the model: it was in the system!

7

CASE 2 OF 3

AI Medical Diagnosis

Google Health — Diabetic Retinopathy

- High diagnostic accuracy in lab settings
- In deployment, nurses had to upload high-quality images manually
- When image quality was insufficient, patients were turned away or misdiagnosed

Epic Sepsis Model

- Poorly calibrated to individual hospital populations
- Integration into EHR systems generated high false positive rates, overwhelming clinicians: real cases were missed

IBM Watson for Oncology

- Poor localization, overreliance on synthetic cases, disconnect from clinical workflows
- Harmful treatment suggestions, including unsafe chemotherapy regimens

Pattern: Failures in data pipelines, infrastructure integration, and clinical oversight.
Even accurate models become ineffective, or harmful, in broken systems.

8

CASE 3 OF 3

High-accuracy perception does not guarantee safe autonomous behavior

Autonomous Vehicles

AV trustworthiness depends on:

- Sensor fusion and decision policies
- Human–machine interaction design
- Fail-safe mechanisms and subsystem communication

Waymo Routing Bug

- Perception correctly identified road cones and traffic
- A failure between AI perception and route planning caused indecision
- The vehicle looped indefinitely

Tesla Sudden Acceleration Claims

- No defects were found in the AI systems
- The lack of robust input validation weakened system reliability
- Absence of human-machine interaction safeguards affected perceived trustworthiness
- This demonstrates how system-level design influences user trust

Pattern: Even accurate perception cannot guarantee safe behavior without reliable system-level coordination. These cases do not diminish the importance of AI-centric research: they show that system-level considerations must complement model-level guarantees.

9

SUMMARY

Where Did Each System Break?

System Layer	Amazon Hiring	Google Health	Waymo AV	Tesla AV
AI/ML model	✓	✓	✓	✓
Data pipeline	✗	✓	—	—
Infrastructure	—	✗	✓	—
Human–system interaction	✓	✓	✓	✗
Governance & oversight	✓	✓	—	✓

✓ = No significant failure at this layer ✗ = Documented failure — = Not applicable / not reported

Key finding: Human-system interaction and governance failed in ALL four cases. The AI/ML model itself was not the primary failure point in any of them.

10

WHY AI-CENTRIC

APPROACHES FALL SHORT

AI does not control its ecosystem

Trust is a property of the entire system

Emergent behavior arises from interactions

11

ARGUMENT 1 OF 3

AI Does Not Control Its Ecosystem

AI models depend on the data they receive, the infrastructure they operate in, and the mechanisms through which their decisions are implemented

Predictive Policing

Fairness-constrained models can still reinforce biases when fine-tuned on historically biased crime data, regardless of their internal fairness mechanisms

Medical AI Diagnosis

A highly accurate diagnostic model is ineffective if outdated electronic health records lead to incorrect treatment recommendations

The Implication:

- Even the most robust and explainable AI model fails if deployed in an insecure or poorly governed system
- Trustworthiness depends on the integrity of the entire ecosystem: accurate and unbiased data pipelines, reliable infrastructure, and effective governance and oversight mechanisms

12

ARGUMENT 2 OF 3

Trust Is a System Property

Trust is often mistakenly attributed to the AI model; in reality, it emerges from the entire system

Data Integrity

AI models receive accurate, unbiased, well-maintained data

Infrastructure

Consistent operation under varying conditions, protected against attacks

Governance

Clear accountability mechanisms and regulatory adherence

Human Oversight

AI outputs reviewed, validated, and responsibly acted upon

Concrete Illustration: AI-powered loan approval system

- The model meets all fairness and explainability standards, still the system as a whole is untrustworthy if the banking infrastructure lacks transparency
- Decision makers fail to interpret AI recommendations correctly, or regulatory safeguards are inadequate

13

ARGUMENT 3 OF 3

Emergent Behavior from Interactions

Undesired behavior often arises from complex interactions among components, not from any single element, even if each part functions correctly in isolation

Model/Data Interaction

- Well-trained models produce flawed outputs when data distributions shift over time
- Produce flawed outputs when exposed to subtle biases in the data

Cross-Component Interactions

- Adaptive models influence user behavior, which alters input data
- This creates self-reinforcing cycles with unintended effects

Human/AI Dynamics

- Users overtrust or undertrust AI based on presentation or perceived authority
- This results in poor decisions or system misuse

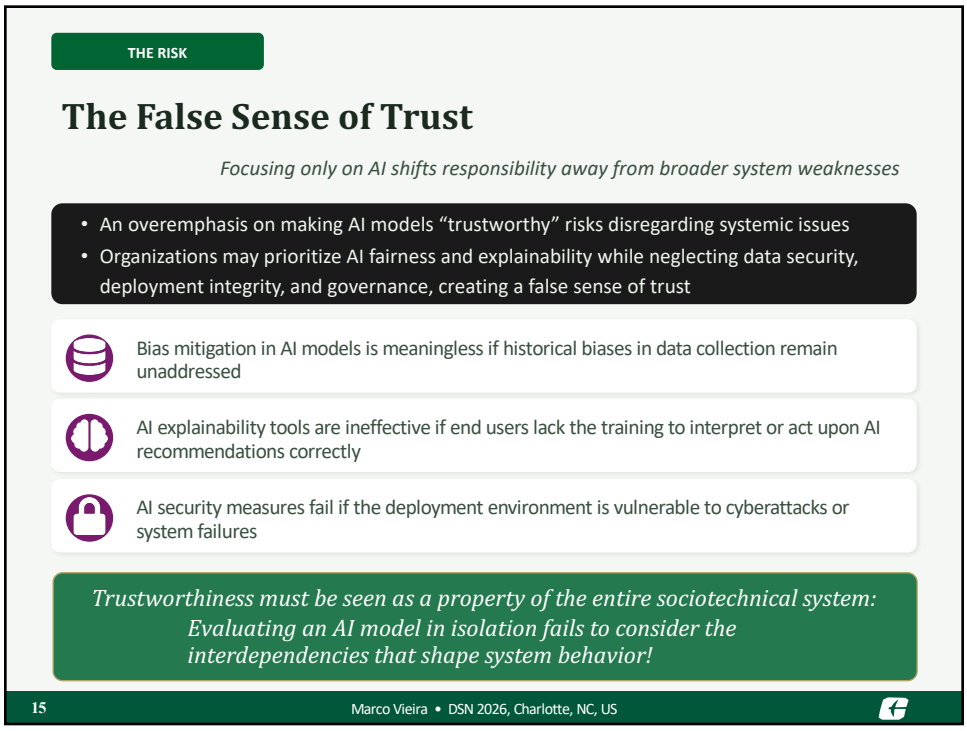
Policy/Technology Misalignment

- Governance frameworks fail to account for real-world deployment complexity
- As a result, systems violate legal or ethical norms despite being technically compliant

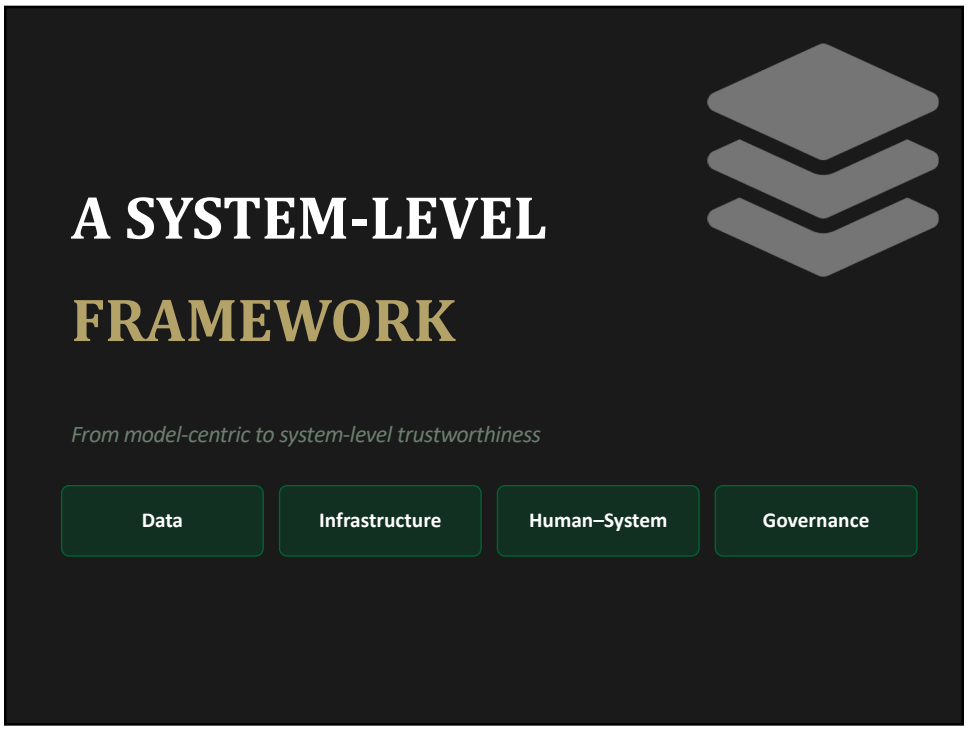
Example:

- Individually reliable content moderation components may interact in unexpected ways
- Feedback loops between components can amplify misinformation
- These interactions can also suppress legitimate content

14


THE RISK
The False Sense of Trust
Focusing only on AI shifts responsibility away from broader system weaknesses
• An overemphasis on making AI models "trustworthy" risks disregarding systemic issues
• Organizations may prioritize AI fairness and explainability while neglecting data security, deployment integrity, and governance, creating a false sense of trust
Bias mitigation in AI models is meaningless if historical biases in data collection remain unaddressed
AI explainability tools are ineffective if end users lack the training to interpret or act upon AI recommendations correctly
AI security measures fail if the deployment environment is vulnerable to cyberattacks or system failures
Trustworthiness must be seen as a property of the entire sociotechnical system: Evaluating an AI model in isolation fails to consider the interdependencies that shape system behavior!
15
Marco Vieira • DSN 2026, Charlotte, NC, US

15


A SYSTEM-LEVEL
FRAMEWORK
From model-centric to system-level trustworthiness
Data
Infrastructure
Human–System
Governance

16

FRAMEWORK

The Multilayered Trustworthiness Model

Layer 4: Governance & Compliance
Ethical principles, legal standards, accountability mechanisms, regulatory adherence (GDPR, EU AI Act)

Layer 3: Human–System Interaction
Human-in-the-loop mechanisms, interpretable interfaces, feedback channels, oversight and accountability

Layer 2: Infrastructure
Reliable computing environments, secure deployment, fault-tolerant operations, protected against attacks

Layer 1: Data
Accurate and unbiased data pipelines, data provenance tracking, drift detection, manipulation prevention

AI model sits inside Layer 1: Trustworthiness is assessed across all four layers in combination

17

FRAMEWORK

Trustworthiness Is Dynamic, Not Static

A model considered fair at training may become unreliable if:

- Deployed on outdated or shifted data
- Operated in a poorly monitored environment
- Subject to regulatory changes not reflected in its governance

Trustworthiness extends throughout the entire AI lifecycle:

Design | Training | Testing | Deployment | Monitoring | Decommissioning

Our system-level perspective is consistent with and extends existing frameworks:

NIST AI RMF
Lifecycle-based assurance spanning design-time, operation-time, and contingency management

TrustOps
Continuous evidence collection methodology to establish a robust trust model

U.S. DoT AI Assurance
Design-time + operation-time + contingency management for transportation systems

18

OPERATIONALIZING

SYSTEM-LEVEL TRUST

Empirical methods from the dependability community

- Fault Injection
- Robustness Testing
- Failure Prediction

19

OPERATIONALIZATION

Assessing Data Trustworthiness

AI models are only as trustworthy as the data they consume

Key Challenges

- Biased training data
- Data drift over time
- Labeling errors
- Data poisoning attacks

Assessment Approaches

- Bias audits
- Statistical drift detection
- Data provenance tracking
- Manipulation detection

20

OPERATIONALIZATION

Fault Injection for Infrastructure Assessment

Deliberately introducing faults to observe system behavior under stress and characterize its resilience

Hardware Faults
Memory errors, disk failures, CPU corruption

Network Disruptions
Packet loss, latency spikes, partition events

Software Faults
Code mutations, configuration errors, runtime bugs

Adversarial Inputs
Corrupted sensor data, poisoned API responses

21

OPERATIONALIZATION

Robustness Testing & Failure Prediction

Robustness Testing

Evaluating system behavior under stress, degraded inputs, and adversarial conditions to characterize failure modes and inform resilience improvements

- How does the system degrade gracefully under load?
- Does the AI component fail safely when infrastructure degrades?
- Are failure recovery mechanisms effective?

Failure Prediction

- Proactive failure detection is essential for dependable autonomous systems
- Online failure prediction uses ML models trained on system telemetry to anticipate failures

Key challenge:

- Failure prediction data generated must be representative
- Otherwise, the predictor will not generalize

22

OPERATIONALIZATION

Human–System & Governance Trustworthiness

Human–System Layer

The interaction layer is where trust is ultimately won or lost:

- Human-in-the-loop mechanisms keeping humans meaningfully in control
- Clear interfaces that surface AI uncertainty and confidence levels
- Feedback channels that allow correction and oversight

Governance & Legal Layer

Alignment with laws, moral principles, and societal expectations:

- GDPR and data protection regulations
- Anti-discrimination laws and the EU AI Act
- Transparency in system behavior
- Accountability mechanisms for oversight and remediation

Focusing only on AI shifts responsibility away from broader system weaknesses. Trust requires coordinated governance, not only technical solutions

23

FRAMEWORK

Context-Aware Prioritization

Trust is not universal: it must be tailored to purpose, risk profile, and expectations

High-stakes: Medical Diagnostic AI

Demands stronger guarantees for interpretability, robustness, and regulatory compliance

Low-stakes: Recommender System

Lower risk threshold allows different trade-offs between performance and explainability

Trustworthiness properties may exist in tension:

- Privacy through data minimization → reduces auditability
- Increasing model interpretability → may cost accuracy or scalability
- Security hardening → may reduce performance

Future work: multi-objective optimization frameworks and stakeholder-informed prioritization strategies

24

FRAMEWORK

Aligning with Existing Standards

NIST AI RMF
Lifecycle-based assurance: needs extension to capture system-level interactions

ISO/IEC 25010
Quality models covering reliability, security, maintainability: needs AI-specific extension

Confiance.ai
Trustworthiness in high-stakes domains (healthcare, defense) with measurable KPIs

Mapping existing standards to the multilayered model will identify where systemic interactions remain underexplored

25

OPEN CHALLENGES

"Trust but verify, continuously": the operational mindset for public-facing AI systems

Open Challenges

System-Level Assessment Frameworks
Evaluation methodologies encompassing data integrity, infrastructure dependability, human–AI interaction, and governance transparency

Trustworthiness Maturity Models
Capability models guiding organizations to assess and progressively improve their AI ecosystems — inspired by CMMI

Assurance Case Approaches
Adapting structured assurance practices from safety-critical domains to AI-powered systems, enabling evidence-based arguments for system-level trust

Risk Propagation Modeling
Techniques to analyze how vulnerabilities in one subsystem propagate through the system and affect overall trust

Context-Aware Prioritization
Methods for balancing competing trustworthiness goals, including multi-objective optimization and stakeholder-informed strategies

Empirical Case Studies
Systematic evidence to validate the framework in real-world system behavior — going beyond illustrative examples

26

CALL TO ACTION

A Call to the Dependability Community

Ongoing work in fairness, robustness, and explainability is essential

However, existing efforts are insufficient if the surrounding data pipelines, infrastructure, governance, and oversight mechanisms remain unreliable

The dependability community has the tools and methods to lead this effort:

- Fault Injection
- Robustness Benchmarking
- Failure Prediction
- Resilience Assessment
- System Testing
- Verification & Validation

Interdisciplinary research is crucial to bridge gaps between software engineering, human–computer interaction, cybersecurity, ethics, and policy

- *Trust must be cultivated at the system level, not just within the AI component*
- *Operationalize system-level trustworthiness for AI-enabled autonomous systems*

27

Thank you!

Marco Vieira

marco.vieira@charlotte.edu

https://marcovieira.me

Trusting the System, Not Just the Model:

A Perspective on AI-Enabled Autonomous Systems

- AI trust is a system property, not a model property
- Four layers: Data • Infrastructure • Human–System • Governance
- Empirical methods from dependability can operationalize system-level trust

28