

REVIEW

Open access

Can Foundation Models Be Trusted? A Scoping Review of Interpretability, Hallucination, Uncertainty, Alignment, Evaluation, and Human Oversight

David Chen^{1*}, Maria Papadopoulou², Kwame Boateng³

Abstract

Foundation models have expanded the scope of artificial intelligence by supporting language generation, multimodal reasoning, knowledge-intensive assistance, decision support, and adaptable downstream applications. Their broad capabilities, however, do not establish that their outputs, explanations, internal processes, or deployment arrangements are trustworthy. This scoping review maps the research landscape concerning foundation-model trustworthiness across interpretability, explanation capabilities, hallucination, uncertainty, reliability, alignment, evaluation, and human oversight. The review uses structured search concepts, eligibility principles, evidence extraction, and dimension-based mapping to examine how peer-reviewed scholarship characterizes foundation models as general-purpose intelligent systems and how trustworthiness evidence is distributed across technical, human-centered, organizational, and governance domains. The synthesis indicates that interpretability methods can expose selected behavioral or representational features but do not independently demonstrate explanation fidelity, usefulness, or safety. Hallucination and factuality research identifies persistent risks involving fabrication, unsupported claims, weak grounding, and persuasive misinformation, while uncertainty methods remain constrained by calibration, context shift, and user-communication challenges. Alignment and post-training techniques may shape preferred behaviors but retain risks associated with proxy objectives, preference representation, and residual failure. Benchmarking, human evaluation, stress testing, and oversight provide partial evidence, yet their value depends on construct validity, domain relevance, intervention capacity, accountability, and lifecycle monitoring. The research landscape is therefore fragmented across disciplines and frequently model-centred rather than socio-technical. Key priorities include stronger foundation-model interpretability validation, consistent hallucination measurement, usable uncertainty communication, externally valid evaluation, meaningful oversight testing, multimodal analysis, and lifecycle governance. The review contributes an integrated evidence map while emphasizing that no single trustworthiness dimension certifies safety, reliability, accountability, or deployment readiness.

Keywords Interpretability, Trustworthy AI, Large language models, Uncertainty, Foundation models, Hallucination

*Correspondence:

David Chen

david.chen@nus.edu.sg

¹ Department of AI and Foundation Models, National University of Singapore, Singapore, Singapore

² Department of Informatics, Athens University of Economics and Business, Athens, Greece

³ Department of Computer Science, University of Ghana, Accra, Ghana

Introduction

Foundation-model trustworthiness has become a central problem for trustworthy and explainable intelligent systems because broad capability, linguistic fluency, and adaptable task performance can coexist with opacity, factual error,

uncertainty, behavioral instability, and weak accountability. Trustworthiness is consequently better understood as a multidimensional lifecycle property requiring coordinated technical, human, ethical, and governance evidence rather than capability or benchmark performance alone [1–3]. This distinction is especially important for models that can be reused across domains, combined with external tools, embedded in decision-support workflows, or adapted for consequential applications without being redesigned from first principles.

Trust must also be distinguished from trustworthiness. Human users may develop confidence in an artificial intelligence system because its responses are fluent, responsive, personalized, or presented with apparent authority, even when its actual reliability has not been established. Empirical research on human–AI interaction shows that trust is affected by observable performance, transparency cues, system appearance, task characteristics, and user expectations [4]. A trustworthy foundation model therefore cannot be defined by whether users believe it, prefer it, or find it understandable. The relevant question is whether reliance is justified by evidence concerning the model, the surrounding system, the intended context, and the consequences of failure.

Existing scholarship addresses parts of this problem through partially separated literatures. Explainable-AI research examines interpretability, post-hoc explanations, fidelity, and human understanding; natural-language-generation research studies hallucination and factuality; machine-learning research develops uncertainty and calibration methods; alignment research examines preference learning and human feedback; evaluation research focuses on benchmarks and human judgments; and human-factors scholarship investigates supervision, intervention, and automation bias. Governance and accountability research adds further requirements concerning responsibility, documentation, auditability, monitoring, and redress. Studying these areas independently can obscure interactions among them, such as when an apparently useful explanation increases reliance on an unreliable output or when strong benchmark performance conceals poor behavior under context shift.

This article therefore provides an original scoping review of whether and under what evidential conditions foundation models may be considered trustworthy. Its objective is not to produce a universal trust score, certify particular models, or prescribe a deployment-ready assurance framework. Instead, it maps how the selected peer-reviewed literature conceptualizes foundation models as general-purpose intelligent systems; charts evidence concerning interpretability, hallucination, uncertainty, reliability, alignment, evaluation, and human oversight; identifies cross-dimensional patterns and tensions; and clarifies areas in which evidence remains limited. The resulting contribution is an integrated research-landscape synthesis designed to support cautious evaluation and future inquiry rather than categorical judgments that foundation models can or cannot be trusted.

Foundation Models as General-Purpose Intelligent Systems

Foundation models are distinguished by their capacity to support multiple downstream tasks through large-scale pretraining, adaptation, prompting, fine-tuning, or integration with specialized components. Their general-purpose character allows one model family to contribute to text production, classification, question answering, summarization, image interpretation, multimodal reasoning, and domain-specific assistance. Research on generalist medical artificial intelligence illustrates both the potential breadth of foundation-model adaptation and the continuing need for task-specific, population-specific, and workflow-specific validation [5]. Generality should therefore be treated as a capability characteristic, not as evidence that performance, safety, or reliability transfers automatically across contexts.

Generative AI broadens this challenge by producing new linguistic, visual, auditory, or structured content rather than merely selecting among predefined outputs. Such systems may enhance creativity, information access, workflow support, and interaction flexibility, but they also introduce risks involving fabricated content, hidden dependencies, intellectual-property concerns, privacy, misuse, and uncertain organizational responsibility [6]. “Generative AI” is consequently not synonymous with “foundation model,” although the categories overlap substantially. A foundation model describes a broadly reusable pretrained model, whereas generative AI emphasizes the production of content; large language models and large multimodal models represent important, but not exhaustive, technical forms within this wider landscape.

Foundation models must additionally be understood as components of socio-technical systems. Their behavior is influenced not only by architecture and training data but also by computing infrastructure, provider policies, interfaces, access arrangements, organizational incentives, deployment choices, and the distribution of expertise and decision authority. Analyses of the political economy of large language models show that computational concentration, platform control, and infrastructure ownership shape who can develop, inspect, modify, and govern these systems [7]. Trustworthiness evidence must therefore extend beyond isolated model outputs to the institutions and technical arrangements through which foundation models are created and used.

Large language models themselves comprise heterogeneous architectures, training procedures, adaptation strategies, retrieval mechanisms, safety layers, and evaluation practices. Technical surveys describe substantial variation in model development and application while identifying persistent challenges concerning efficiency, factuality, robustness, reasoning, privacy, evaluation, and responsible deployment [8]. Capability scale can increase the range of tasks a system performs, but it also expands the surface on which failures can occur. Similarly, fluency can make outputs more accessible while making unsupported claims more persuasive. Foundation models should therefore be evaluated as general-purpose but bounded intelligent systems whose trustworthiness remains dependent on intended use, evidence quality, human interaction, organizational controls, and continuing monitoring.

Scoping Review Questions and Evidence-Mapping Method

The review was organized around six scoping review questions: how foundation models are characterized as general-purpose intelligent systems; what evidence exists on interpretability and explanation capabilities; how hallucination, uncertainty, reliability, robustness, and factuality are conceptualized and evaluated; how alignment, benchmarking, human evaluation, and oversight are connected to trustworthiness; what landscape patterns emerge across these dimensions; and what evidence gaps remain. Reporting was guided by established scoping-review principles emphasizing transparent questions, eligibility decisions, evidence charting, and synthesis, while avoiding unsupported claims of exhaustive systematic coverage or quantitative aggregation [9].

A scoping approach was selected because the relevant evidence is conceptually broad, methodologically heterogeneous, and distributed across technical, behavioral, human-factors, organizational, and governance research. Scoping reviews are suited to clarifying key concepts, identifying the types and distribution of available evidence, mapping research approaches, and locating underdeveloped areas rather than estimating a pooled intervention effect [10]. The unit of synthesis was therefore not a standardized outcome measure. Instead, articles were mapped according to the foundation-model context addressed, the trustworthiness dimensions examined, the type of evidence supplied, the limitations reported, and the precise manuscript claim that the evidence could support.

Search concepts combined terms for foundation models, large language models, large multimodal models, generative AI, general-purpose AI, trustworthy AI, explainable AI, interpretability, hallucination, factuality, uncertainty, calibration, reliability, robustness, alignment, reinforcement learning from human feedback, evaluation, benchmarking, human oversight, accountability, and lifecycle governance. Eligibility required a peer-reviewed journal article, verifiable DOI, suitable journal quality, substantive relevance to one or more review questions, and sufficient information for evidence extraction. Methodological guidance for scoping reviews informed the use of explicit eligibility principles, source-selection logic, charting categories, and a transparent relationship between the questions and synthesis structure [11].

Evidence was charted across model context, interpretability or explanation features, hallucination and uncertainty risks, alignment and evaluation functions, human-oversight arrangements, governance implications, study limitations, and research-gap relevance. Where an article did not clearly address a category, the category was treated as unreported rather than inferred. Extraction and presentation were aligned with the review questions so that the synthesis would represent evidence distribution and conceptual relationships rather than numerical prevalence or comparative effectiveness [12]. **Table 1** summarises the scoping review logic used to identify, screen, map, and synthesize evidence

on foundation model trustworthiness. **Figure 1** illustrates the scoping review evidence-mapping workflow used to move from review questions and search concepts to foundation-model relevance screening, trustworthiness-dimension mapping, synthesis, and gap identification.

Table 1. Scoping Review Logic for Foundation Model Trustworthiness: Review Questions, Search Concepts, Inclusion Criteria, Exclusion Criteria, Screening Logic, Evidence-Mapping Categories, Foundation Model Relevance, Trustworthiness-Dimension Mapping, Citation Planning, and Review Boundaries

Review component	Core review question or purpose	Inclusion logic	Exclusion logic	Evidence mapped	Trustworthiness dimension addressed	Bias or limitation risk
Review questions	Define the scope of the evidence map	Addresses at least one stated scoping question	No connection to the review objectives	Concepts, methods, findings, limitations	Cross-cutting	Questions may privilege established dimensions
Search concepts	Capture interdisciplinary terminology	Uses direct or closely related foundation-model and trustworthiness concepts	Narrow performance-only terminology	Model classes, trust concepts, risks, controls	All dimensions	Terminological variation
Journal eligibility	Establish peer-reviewed evidence quality	Peer-reviewed journal article with verifiable DOI and suitable journal standing	Preprints, conferences, websites, reports, or missing DOI	Bibliographic and publication status	Documentation	Quartiles vary by year and category
Foundation-model relevance	Connect evidence to the review object	Direct FM, LLM, multimodal, generative-AI evidence or justified transferable evidence	No defensible general-purpose-AI relevance	Model context and transferability	Foundation model; general-purpose AI	Generic AI evidence may be overextended
Trustworthiness mapping	Classify evidence by substantive dimension	Supports one or more predefined trustworthiness categories	Superficial mention without usable evidence	Interpretability, reliability, alignment, oversight, governance	Dimension-specific and cross-cutting	Better-developed topics may dominate
Title and abstract screening	Remove clearly irrelevant records	Plausible conceptual or empirical fit	Narrow application with no transferable trust relevance	Stated aim and scope	Initial dimension relevance	Abstracts may omit useful evidence

Full-text assessment	Confirm exact claim support	Full text supports an extractable claim or identified gap	Weak or indirect fit that cannot support planned wording	Methods, findings, limitations	Claim-specific	Confirmation bias
Evidence extraction	Chart comparable information	Reports sufficient contextual and substantive detail	Evidence cannot be reliably extracted	Study aim, context, features, limitations	All dimensions	Heterogeneous reporting
Citation planning	Fix first appearance and sentence-level fit	Supports a specific planned manuscript sentence	Topic-only or random citation placement	Claim–source relationship	Citation integrity	Citation clustering
Synthesis boundaries	Prevent overclaiming	Supports cautious mapping and gap identification	Claims of certification, universal metrics, or guaranteed readiness	Evidence patterns, tensions, gaps	No-sufficiency boundaries	Scoping evidence cannot prove effectiveness

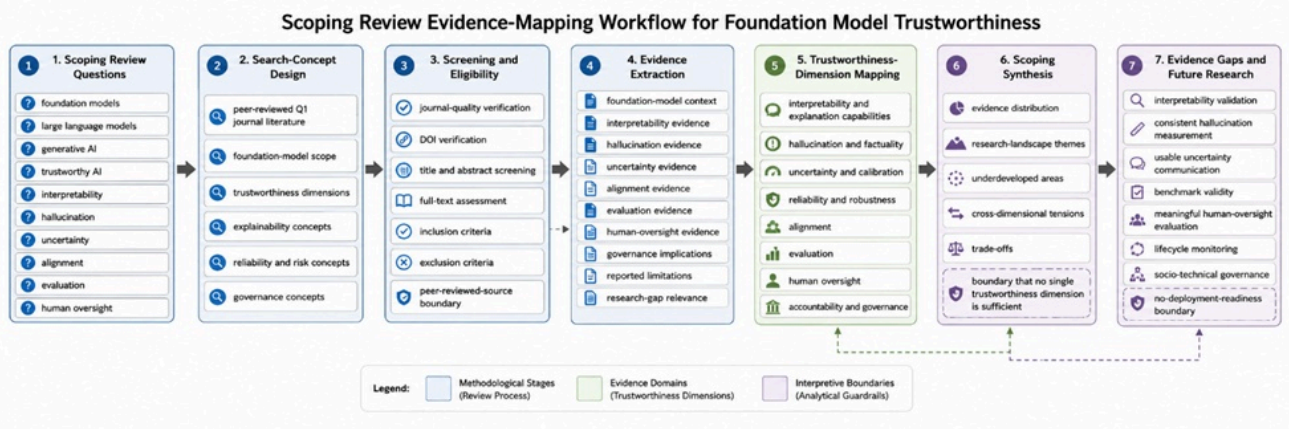


Figure 1. Scoping Review Evidence-Mapping Workflow for Foundation Model Trustworthiness

Alt-text description. A scoping review workflow diagram showing review questions, search concepts, peer-reviewed journal screening, eligibility checks, foundation-model relevance mapping, trustworthiness-dimension mapping, evidence extraction, scoping synthesis, and evidence-gap identification.

Interpretability and Explanation Capabilities

Interpretability and explanation are related but non-identical concepts. Interpretability generally concerns the extent to which relevant model behavior, internal representations, or input–output relationships can be understood, whereas explanations are representations communicated for a particular purpose and audience. Interpretability and explanation

comprise heterogeneous technical methods and human-facing practices whose value depends on context, purpose, and responsible-AI requirements [13-15]. For foundation models, these practices may include internal probing, representation analysis, feature attribution, concept-based analysis, counterfactual reasoning, prompt-based diagnostics, output rationales, source-linked explanations, and natural-language accounts. No single method reveals the model in its entirety.

Post-hoc explanations are particularly important because many foundation models are too large, distributed, or proprietary for complete inspection. However, an explanation generated after an output may describe an associated pattern without faithfully representing the process that produced the output. Explanation quality must therefore be separated into dimensions such as fidelity, stability, relevance, completeness, uncertainty, and usefulness. Reviews of explainable AI show that methods, datasets, evaluation metrics, and intended users remain heterogeneous, and that explanation availability alone does not establish trustworthy behavior [16]. A fluent rationale may be plausible yet unfaithful, while a technically faithful representation may remain unusable to a non-specialist.

Mechanistic interpretability seeks stronger causal understanding by examining internal components, activations, representations, or computational pathways associated with model behavior. Where supported, causal interventions may provide more informative evidence than observational correlations. Nevertheless, journal evidence specific to large-scale foundation models remains uneven, particularly across proprietary systems, model versions, modalities, and deployment contexts. Recent research on vision foundation models demonstrates expanding interest in attribution, prompting, concept-based analysis, and visual explanation, while also identifying substantial evaluation gaps [17]. Multimodal models create additional challenges because explanations may need to account for relationships among text, images, audio, sensor inputs, and generated outputs.

Human-centred explanation evaluation asks whether explanations improve comprehension, support error detection, enable contestation, or calibrate reliance. These outcomes cannot be inferred from technical metrics alone. Natural-language explanations may improve cognitive accessibility but can also conceal uncertainty through persuasive presentation. Feature-attribution methods may identify influential inputs without demonstrating causality, while counterfactual explanations may be invalid, infeasible, or unstable. Evaluation should therefore be matched to the stakeholder: developers may require debugging evidence, auditors may need traceability and reproducibility, domain professionals may need source grounding and uncertainty qualification, and affected users may need understandable reasons, alternatives, and avenues for challenge.

Interpretability should ultimately be treated as one contributor to a broader trustworthiness assessment. An interpretable system may still generate harmful, biased, insecure, or factually incorrect outputs, and an apparently understandable explanation may not establish alignment, accountability, or safe deployment. Foundation-model explanations require explicit scope statements, fidelity assessment, stability testing, uncertainty communication, and human-centred validation. The no-interpretability-sufficiency boundary therefore prevents explanation evidence from being used as a substitute for reliability, alignment, oversight, or governance evidence. **Table 2** maps interpretability and explanation capabilities relevant to trustworthy foundation models, including technical functions, stakeholder needs, failure risks, and evidence gaps.

Table 2. Interpretability and Explanation Capabilities in Foundation Models: Mechanistic Interpretability, Model Transparency, Post-Hoc Explanation, Explanation Quality, Usefulness, Fidelity, Stability, Uncertainty, Feature Attribution, Counterfactuals, Natural-Language Explanations, Human Understanding, Cognitive Accessibility, and Explanation Boundaries

Interpretability or explanation capability	Core trustworthiness question	Technical or human-centered function	Potential evaluation approach	Primary stakeholder	Failure or misuse risk	Evidence gap

Interpretability	Can relevant behavior be understood beyond output plausibility?	Exposes model patterns or decision relationships	Behavioral probes, expert inspection, causal testing	Researchers, developers, auditors	Partial understanding may be presented as complete	Limited validation at foundation-model scale
Mechanistic interpretability	Do identified internal components causally influence behavior?	Connects representations or circuits to outputs	Intervention, ablation, activation patching, replication	Safety researchers and evaluators	Correlation may be mistaken for mechanism	Sparse cross-model journal evidence
Model transparency	Are development, data, architecture, limitations, and changes documented?	Supports scrutiny and traceability	Documentation and disclosure audit	Auditors, deployers, researchers	Selective disclosure may create false confidence	Restricted access to proprietary systems
Post-hoc explanation	Can a comprehensible account be produced after an output?	Provides local or global explanation	Fidelity, perturbation, sensitivity, and user testing	Users, reviewers, developers	Rationalization may not reflect actual processing	Limited FM-specific validation
Explanation quality	Is the explanation relevant, accurate, appropriately scoped, and complete enough?	Evaluates multiple explanation properties	Multi-criteria rubric and expert review	All stakeholders	Polished language may mask weak evidence	No universal quality standard
Explanation usefulness	Does the explanation improve decisions or error detection?	Supports practical action and comprehension	Task performance, error recognition, reliance calibration	End users and domain experts	Explanations may increase workload or overreliance	Few longitudinal high-stakes studies
Explanation fidelity	Does the explanation correspond to actual model behavior?	Tests faithfulness	Causal intervention, perturbation, deletion, consistency	Researchers and auditors	Fluent but unfaithful rationales	Difficult validation for natural-language explanations

Explanation stability	Are explanations consistent under immaterial changes?	Tests explanatory robustness	Paraphrase, seed, prompt, and version sensitivity	Developers and auditors	Unstable accounts weaken accountability	Limited longitudinal and cross-version evidence
Explanation uncertainty	Does the system qualify uncertainty in its explanation?	Communicates evidential limits	Alternative explanations, confidence validation, abstention	Domain professionals and users	Definitive wording may hide ambiguity	Rarely measured directly
Feature attribution	Which inputs are associated with the output?	Produces importance scores or maps	Sanity checks, perturbation, robustness analysis	Developers and auditors	Attribution may be method-dependent or non-causal	Weak semantic validation for generative outputs
Counterfactual explanation	What change could alter the model output?	Supports contrastive understanding and recourse	Validity, feasibility, sparsity, stability	Affected users and developers	Counterfactuals may be unrealistic or manipulable	Limited evidence for open-ended generation
Natural-language explanation	Can reasons and limitations be communicated accessibly?	Provides user-facing accounts	Fact verification, grounding, fidelity, comprehension tests	General users and professionals	Eloquence may be confused with truth	No reliable self-explanation guarantee
Human understanding	Does explanation support a correct mental model?	Improves comprehension and error recognition	Mental-model and prediction-of-behavior tests	Users and supervisors	Users may misunderstand technical qualifiers	Limited stakeholder diversity
Cognitive accessibility	Is the explanation usable across expertise and cognitive needs?	Adapts format, terminology, and detail	Accessibility, comprehension, and workload evaluation	Diverse user groups	Oversimplification may conceal uncertainty	Sparse accessibility-centred evidence
No interpretability-sufficiency boundary	Does interpretability independently prove trustworthiness?	Prevents substitution for other evidence	Cross-dimensional assessment	Evaluators and deployers	Visibility may be mistaken for safety	Few integrated assurance studies
No explanation-	Does an explanation prove	Separates communication	Fidelity, usefulness, stability, and	All stakeholders	Explanation presence may	No universal certification method

guarantee boundary	correctness, fidelity, or usefulness?	from validated evidence	uncertainty testing		imply unjustified accountability	
-----------------------	---	----------------------------	------------------------	--	-------------------------------------	--

Hallucination, Uncertainty, and Reliability Risks

Hallucination is a central foundation-model reliability problem because generated content may be syntactically coherent and contextually plausible while unsupported by available evidence. The literature uses the term to cover several related but distinguishable failures, including fabrication, contradiction, unsupported inference, incorrect attribution, and divergence from source material. These failures vary across open-ended generation, summarization, dialogue, translation, and knowledge-intensive tasks, making hallucination a family of context-dependent errors rather than a single measurable phenomenon [18].

Large language model research further distinguishes factuality, groundedness, and truthfulness. Factuality concerns whether individual claims are correct, groundedness concerns whether outputs are supported by specified evidence, and truthfulness includes whether the system avoids misleading presentation and communicates relevant limitations. Current evidence remains fragmented across definitions, benchmarks, annotation protocols, detection methods, and mitigation strategies, limiting direct comparison across models and tasks [19]. Fluency may worsen this risk by making false content easier to accept, repeat, or incorporate into downstream decisions.

Uncertainty estimation offers one route for identifying unreliable outputs, but it does not automatically verify factual correctness. Semantic entropy, for example, evaluates variation in meaning across sampled generations and may help distinguish stable responses from outputs reflecting unresolved knowledge uncertainty [20]. Such methods are promising for selective prediction, abstention, or escalation, yet their effectiveness depends on task design, sampling assumptions, model access, and the relationship between semantic variation and the relevant error type. A confident or semantically stable response may still be wrong because the model consistently reproduces an incorrect representation.

Uncertainty must also be interpreted according to its source. Epistemic uncertainty concerns incomplete knowledge and may sometimes be reduced through additional evidence, improved training, retrieval, or expert review. Aleatoric uncertainty concerns ambiguity or variability that cannot necessarily be removed by obtaining more data. Deep-learning research offers diverse approaches to quantification, confidence estimation, calibration, and out-of-distribution detection, but their assumptions and validity differ across architectures and use contexts [21-23]. A numerical confidence score is therefore not inherently meaningful unless it is calibrated, validated under relevant conditions, and communicated in a form users can interpret appropriately.

Reliability and robustness extend beyond isolated hallucinations. Reliability concerns whether performance remains acceptably consistent for an intended use, whereas robustness concerns resistance to ordinary variation, adversarial manipulation, changing prompts, context shift, and unfamiliar inputs. Safety additionally requires attention to the severity and distribution of consequences, not merely average error. No hallucination-elimination, uncertainty-guarantee, or reliability-guarantee claim is justified by one mitigation or benchmark. Trustworthy use requires layered grounding, verification, uncertainty communication, domain validation, abstention, escalation, and lifecycle monitoring.

Alignment, Evaluation, and Human Oversight

Alignment concerns whether model behaviour corresponds to specified objectives, instructions, preferences, or values. Reinforcement learning from human feedback can shape outputs by learning proxy reward functions from comparative judgments, but its evidential limits include sparse or inconsistent feedback, reward misspecification, unrepresentative

preferences, and incorrect generalization beyond evaluated conditions [24]. Improved instruction following or reduced harmful output on selected tests should therefore be interpreted as evidence of behavioural adaptation rather than proof that the underlying model is comprehensively aligned.

Human and value alignment also raise normative questions about whose goals are represented. Alignment may refer to following immediate instructions, satisfying user preferences, supporting longer-term interests, respecting institutional requirements, or reflecting wider moral values. These targets can conflict, and their selection requires justification rather than purely technical optimization [25]. Instruction tuning, preference learning, safety tuning, and reinforcement learning from human feedback should consequently be evaluated for stakeholder representation, objective transparency, conflict handling, and residual risks.

Evaluation provides evidence about capabilities and failures, but foundation-model evaluation is heterogeneous. Relevant approaches include task benchmarks, factuality testing, calibration assessment, human evaluation, stress testing, adversarial evaluation, red-teaming, out-of-distribution testing, and domain-specific validation. Surveys of large language model evaluation show that methods differ in what they measure, how judgments are obtained, and whether they address capability, robustness, safety, social risk, or user experience [26]. No single evaluation score captures the full trustworthiness profile of a general-purpose model.

Benchmark validity is particularly important because standardized evaluations can become saturated, contaminated, narrowly optimized, or disconnected from real-world requirements. Empirical mapping of artificial-intelligence benchmark ecosystems indicates that benchmark creation and use are concentrated and that performance improvements do not necessarily correspond to durable practical progress [27]. Evaluation should therefore include a clear construct-validity argument, transparent task scope, independent testing, changing-context analysis, and evidence that the evaluated conditions resemble the intended deployment setting.

Human oversight must be designed as an operational control rather than added as a symbolic safeguard. Human-centred AI research emphasizes that high automation can coexist with strong human control when users retain meaningful authority, information, and intervention mechanisms [28]. Human–automation research also shows that supervisors may lose situation awareness, become overloaded, over-rely on system recommendations, or be unable to intervene rapidly enough [29]. Human-in-the-loop approval, human-on-the-loop monitoring, escalation pathways, rollback, and stop controls therefore require direct evaluation of competence, workload, timing, authority, and intervention success. **Table 3** synthesises hallucination, uncertainty, reliability, alignment, evaluation, and human oversight dimensions in foundation model trustworthiness research.

Table 3. Hallucination, Uncertainty, Reliability, Alignment, Evaluation, and Human Oversight in Foundation Model Trustworthiness: Factuality, Groundedness, Truthfulness, Calibration, Confidence, Robustness, Benchmark Validity, Human Evaluation, Red-Teaming, Human Control, Meaningful Oversight, and Sufficiency Boundaries

Trustworthiness dimension	Core risk or evaluation question	Technical or governance function	Potential evidence or evaluation approach	Affected stakeholder or use context	Failure or overclaiming risk	Oversight or governance need
Hallucination and fabrication	Does the system generate unsupported or invented content?	Detects generative error	Claim verification, source comparison, expert review	Information-dependent users	Fluent misinformation	Grounding, abstention, escalation

Factuality and groundedness	Are claims correct and supported by relevant evidence?	Tests correctness and source entailment	Claim-level annotation and evidence tracing	Professional and public users	Unsupported claims presented as sourced	Provenance and independent review	S
Truthfulness and misinformation	Does the system avoid materially misleading communication?	Evaluates behaviour and dissemination risk	Adversarial prompts and scenario-based assessment	Public-facing deployments	Scalable persuasive error	Monitoring and use restrictions	
Uncertainty estimation	Can uncertainty relevant to an output be identified?	Produces uncertainty signals	Semantic entropy, ensemble disagreement, predictive analysis	Developers and experts	Scores mistaken for factual guarantees	Task-specific validation	U
Calibration and confidence	Does expressed confidence correspond to observed correctness?	Assesses confidence reliability	Reliability diagrams and selective prediction	Evaluators and users	Overconfident incorrect output	Recalibration and qualified display	U
Epistemic and aleatoric uncertainty	Is uncertainty reducible through further evidence or intrinsic to the task?	Distinguishes uncertainty sources	Model comparison and ambiguity analysis	Novel or ambiguous contexts	False precision	Retrieval, deferral, alternative presentation	C c
Reliability and robustness	Does behaviour remain stable under variation and shift?	Tests repeatability and resilience	Perturbation, stress, adversarial, and OOD testing	Deployers and affected users	Nominal performance overgeneralized	Revalidation and monitoring	C
Alignment and RLHF	Do post-training objectives produce robust desired behaviour?	Shapes preference-sensitive outputs	Reward-model audits and shift testing	Developers and users	Proxy alignment presented as comprehensive alignment	Objective disclosure and independent assessment	A s
Benchmark validity	Do evaluation tasks measure relevant constructs and contexts?	Supports comparison	Contamination, saturation, construct and external validity checks	Researchers and deployers	High scores treated as readiness	Benchmark governance	E s
Human evaluation	Are human judgments	Assesses nuanced output quality	Rater training, agreement and	Evaluators and affected groups	Bias, fatigue, hidden criteria	Rater documentation	

	reliable and representative?		subgroup analysis				
Red-teaming and stress testing	Can plausible failures and misuse pathways be exposed?	Searches system boundaries	Adaptive adversarial scenarios	Safety and security teams	One test cycle treated as complete	Recurrent testing and transparent scope	
Human control	Can authorised people constrain, stop, or reverse actions?	Maintains decision authority	Override, rollback and intervention tests	Operators and accountable leaders	Controls may be inaccessible or delayed	Tested authority and fallback mechanisms	
Meaningful human oversight	Is review informed, timely, competent, and empowered?	Converts nominal review into effective supervision	Workload, comprehension and intervention assessment	Supervisors and organisations	Human presence treated as a guarantee	Defined roles and escalation	

Scoping Synthesis and Research Landscape

The mapped evidence depicts foundation-model trustworthiness as a socio-technical research landscape rather than a model attribute that can be measured once and generalized across uses. Organizational research describes an automation–augmentation paradox in which AI can substitute for selected human tasks while simultaneously increasing the need for human interpretation, coordination, and judgment [30]. For foundation models, this tension appears in the contrast between expanding general-purpose capability and the growing burden of validation, supervision, and exception handling.

Interpretability research is developing more slowly than capability scale, while explanation availability is advancing faster than validated explanation usefulness. Hallucination research similarly reveals a tension between linguistic fluency and evidential reliability, and uncertainty research shows a gap between technical estimation and user-facing communication. Accountability remains an organizational responsibility even when model processes are opaque, distributed, or difficult to reconstruct [31]. Opacity may complicate investigation, but it does not remove responsibility from providers, integrators, deployers, or decision makers.

Governance research also identifies a persistent gap between principles and operational practice. Reviews of ethics tools find a growing range of assessment methods, documentation templates, checklists, and governance processes, but their maturity, specificity, adoption, and effectiveness vary considerably [32]. Technical evaluation and governance are therefore complementary: model testing cannot determine whether responsibilities are clearly allocated, incidents are reported, affected users can contest decisions, or model changes trigger renewed review.

Deployment readiness requires evidence about the full system lifecycle rather than the model alone. Readiness-level approaches for machine-learning systems emphasize data quality, software integration, human roles, documentation, validation, monitoring, feedback, and change management [33]. Foundation-model trustworthiness should similarly be assessed across development, adaptation, integration, operation, updating, and retirement. Performance may change through new prompts, retrieval sources, tools, user populations, model versions, or environmental conditions even when the underlying model name remains unchanged.

Table 4. Scoping Synthesis and Research Landscape for Trustworthy Foundation Models: Interpretability Versus Scale, Explanation Usefulness, Hallucination and Fluency, Uncertainty Communication, Reliability Under Context Shift, Alignment and Residual Risk, Benchmark Validity, Human Oversight, Accountability, Lifecycle Monitoring, and Future Research Priorities

Research landscape theme	Dimensions involved	Evidence pattern	Dominant research focus	Underdeveloped area	Tension or trade-off	Future research need	
Interpretability versus scale	Interpretability, capability	Capability growth exceeds validated internal understanding	Probing and post-hoc methods	Replicated causal analysis	Generality expands opacity	Cross-model mechanistic validation	U
Explanation availability versus usefulness	Explanation, human understanding	Many methods, limited outcome validation	Explanation generation	Decision usefulness and accessibility	More explanation may increase overreliance	Longitudinal stakeholder studies	
Hallucination versus fluency	Factuality, groundedness	Persuasive language can conceal unsupported claims	Detection and mitigation	Consequence-sensitive validation	Fluency improves usability and misinformation risk	Claim-level verification research	f
Uncertainty estimation versus communication	Calibration, confidence, human factors	Technical methods exceed communication evidence	Uncertainty quantification	User interpretation	Simplification may mislead	Joint technical–human studies	c
Reliability versus flexibility	Robustness, generalization	More use contexts create more failure conditions	Standard evaluations	Context and temporal shift	Flexibility increases validation burden	Prospective OOD evaluation	
Alignment versus residual risk	RLHF, values, safety tuning	Behaviour improves on selected objectives	Post-training methods	Objective and stakeholder transparency	Preference fit may not equal safety or truth	Independent alignment audits	

Benchmarks versus real-world validity	Evaluation, benchmarking	Standardization supports comparison but invites saturation	Leaderboards	External and construct validity	Comparable scores may lack operational meaning	Deployment-linked validation	
Oversight versus automation scale	Human control, HITL, HOTL	Oversight is frequently required but rarely stress-tested	Review interfaces	Workload and intervention performance	Scale can exceed human attention	Incident simulation	D
Accountability versus opacity	Governance, auditability	Opacity complicates but does not remove responsibility	Principles and documentation	Role-specific redress	Responsibility can diffuse across actors	Operational accountability studies	r i
Lifecycle monitoring	Reliability, governance	Predeployment evidence dominates	Initial testing	Drift, updates, incident learning	Static assurance becomes obsolete	Prospective monitoring studies	tr
No single-dimension sufficiency	All dimensions	Each evidence stream remains partial	Dimension-specific metrics	Integrated socio-technical evidence	Improvement in one area may conceal another weakness	Non-compensatory evidence profiles	

Figure 2 presents the foundation model trustworthiness landscape, linking interpretability, hallucination, uncertainty, reliability, alignment, evaluation, human oversight, accountability, governance, and future research gaps.

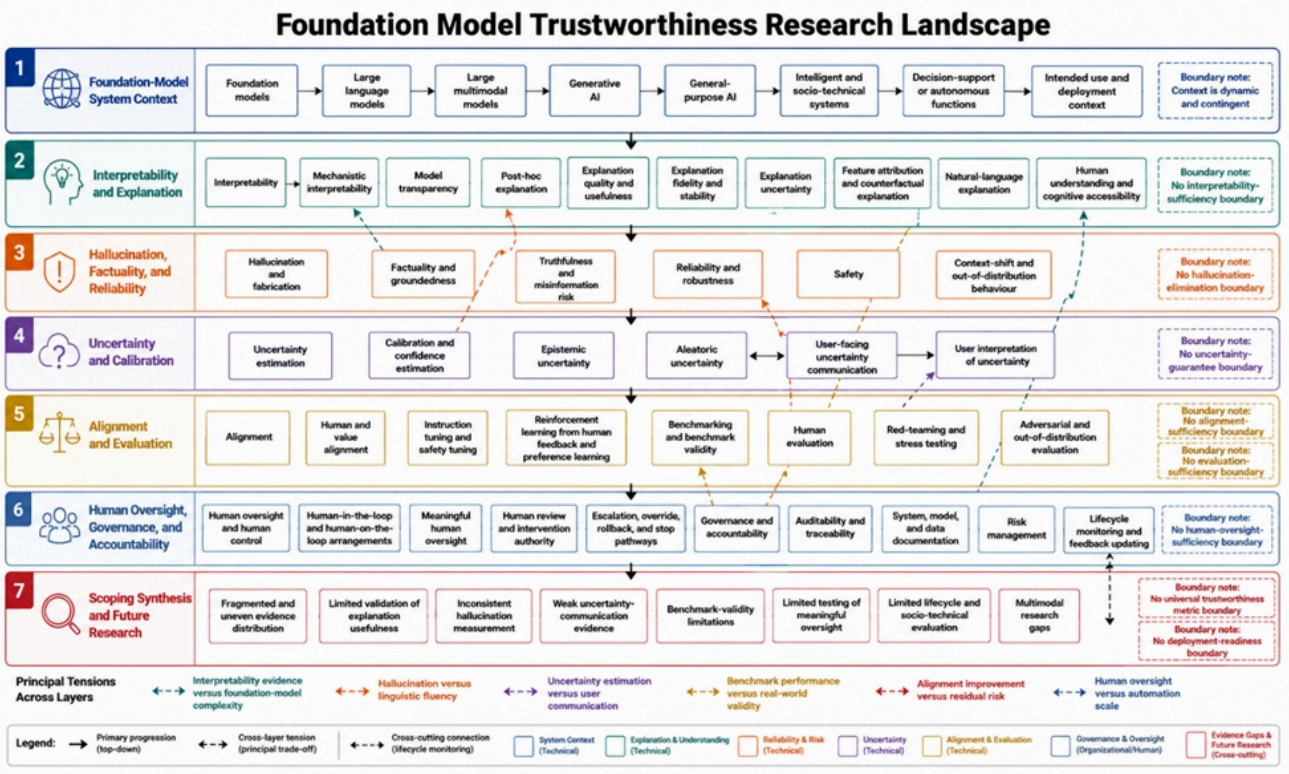


Figure 2. Foundation Model Trustworthiness Research Landscape

Alt-text description

A conceptual research landscape showing foundation model trustworthiness across interpretability, hallucination, uncertainty, reliability, alignment, evaluation, human oversight, accountability, governance, lifecycle monitoring, and evidence gaps.

Evidence Gaps and Future Research Directions

Interpretability evidence remains limited at the scale, opacity, and multimodality of contemporary foundation models. Future research should test whether proposed mechanisms or explanations remain stable across model families, versions, prompts, languages, and domains. Explanation research also requires stronger stakeholder-centred validation. Technical fidelity should be examined alongside usefulness, cognitive accessibility, decision quality, error detection, and reliance calibration rather than treating comprehensible output as sufficient evidence [16].

Hallucination research would benefit from more consistent definitions, claim-level annotation, source-grounding protocols, and evaluation across changing contexts. Uncertainty estimation should be connected to factual verification, abstention, escalation, and user communication. Future studies should examine whether uncertainty displays improve decisions across expertise levels and whether calibration persists under temporal, domain, population, and modality shift. Multimodal systems require particular attention because errors may arise within or between textual, visual, auditory, and sensor representations [19].

Alignment and evaluation research requires greater transparency concerning training objectives, preference sources, annotator populations, reward models, safety trade-offs, and residual failures. Benchmark research should move beyond static leaderboards toward construct validation, contamination assessment, external validation, and longitudinal testing. Red-teaming and adversarial evaluation should report their threat assumptions and limitations because findings may not

generalize across model updates, interfaces, tools, or user populations. Human oversight should be tested through realistic workload, intervention, escalation, and incident scenarios rather than assumed from the presence of a reviewer [29].

The most important cross-cutting gap is limited socio-technical and lifecycle evidence. Research should examine complete systems that include models, retrieval components, tools, interfaces, organizations, users, affected populations, monitoring arrangements, and accountability structures. Prospective studies are needed on drift, model updates, incident learning, feedback incorporation, contestability, and responsibility allocation. Because trustworthiness dimensions are non-substitutable, future evaluation should use transparent evidence profiles and minimum domain-specific requirements rather than universal aggregate scores. These directions are proposed as a research agenda, not as a validated assurance or deployment-readiness framework.

Conclusion

This scoping review maps foundation-model trustworthiness across interpretability, explanation capabilities, hallucination, factuality, uncertainty, reliability, robustness, alignment, evaluation, human oversight, accountability, governance, and lifecycle monitoring. The evidence landscape shows that each dimension provides necessary but partial information: explanations may be unfaithful or unusable, fluent outputs may be false, uncertainty estimates may be miscalibrated or misunderstood, alignment methods may retain proxy and value conflicts, benchmarks may lack real-world validity, and human oversight may be ineffective without authority, information, competence, and time. Foundation models should therefore be evaluated through multidimensional, socio-technical, context-specific, and lifecycle-aware evidence rather than scale, fluency, benchmark performance, alignment claims, explanation availability, or nominal supervision alone. The review contributes an integrated evidence map and identifies priorities for more rigorous interpretability validation, hallucination measurement, uncertainty communication, evaluation validity, oversight testing, multimodal analysis, accountability, and continuous monitoring, while preserving a clear boundary between research synthesis and certification or deployment readiness.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 20 Jul 2025

Revised: 17 Aug 2025

Accepted: 24 Sep 2025

Published online: 15 March 2026

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Thiebes S, Lins S, Sunyaev A. Trustworthy artificial intelligence. *Electron Mark*. 2021;31(2):447-64.
<https://doi.org/10.1007/s12525-020-00441-4>.
- Kaur D, Uslu S, Rittichier KJ, Durresi A. Trustworthy artificial intelligence: a review. *ACM Comput Surv*. 2023;55(2):1-38.
<https://doi.org/10.1145/3491209>.
- Li B, Qi P, Liu B, Di S, Liu J, Pei J, et al. Trustworthy AI: from principles to practices. *ACM Comput Surv*. 2023;55(9):1-46.
<https://doi.org/10.1145/3555803>.
- Glikson E, Woolley AW. Human trust in artificial intelligence: review of empirical research. *Acad Manag Ann*. 2020;14(2):627-60.
<https://doi.org/10.5465/annals.2018.0057>.
- Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. *Nature*. 2023;616(7956):259-65.
<https://doi.org/10.1038/s41586-023-05881-4>.
- Feuerriegel S, Hartmann J, Janiesch C, Zschech P. Generative AI. *Bus Inf Syst Eng*. 2024;66(1):111-26.
<https://doi.org/10.1007/s12599-023-00834-7>.
- Luitse D, Denkena W. The great Transformer: examining the role of large language models in the political economy of AI. *Big Data Soc*. 2021;8(2):20539517211047734.
<https://doi.org/10.1177/20539517211047734>.
- Kumar P. Large language models (LLMs): survey, technical frameworks, and future challenges. *Artif Intell Rev*. 2024;57(10):260.
<https://doi.org/10.1007/s10462-024-10888-y>.
- Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med*. 2018;169(7):467-73.
<https://doi.org/10.7326/M18-0850>.
- Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med Res Methodol*. 2018;18(1):143.
<https://doi.org/10.1186/s12874-018-0611-x>.
- Peters MDJ, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. *JBIEvid Synth*. 2020;18(10):2119-26.
<https://doi.org/10.11124/JBIES-20-00167>.
- Pollock D, Peters MDJ, Khalil H, McInerney P, Alexander L, Tricco AC, et al. Recommendations for the extraction, analysis, and presentation of results in scoping reviews. *JBIEvid Synth*. 2023;21(3):520-32.
<https://doi.org/10.11124/JBIES-22-00123>.
- Miller T. Explanation in artificial intelligence: insights from the social sciences. *Artif Intell*. 2019;267:1-38.
<https://doi.org/10.1016/j.artint.2018.07.007>.

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. *ACM Comput Surv.* 2018;51(5):1-42.
<https://doi.org/10.1145/3236009>.

Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Benetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion.* 2020;58:82-115.
<https://doi.org/10.1016/j.inffus.2019.12.012>.

Ali S, Abuhmed T, El-Sappagh S, Muhammad K, Alonso-Moral JM, Confalonieri R, et al. Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. *Inf Fusion.* 2023;99:101805.
<https://doi.org/10.1016/j.inffus.2023.101805>.

Kazmierczak R, Berthier E, Frehse G, Franchi G. Explainability and vision foundation models: a survey. *Inf Fusion.* 2025;122:103184.
<https://doi.org/10.1016/j.inffus.2025.103184>.

Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. *ACM Comput Surv.* 2023;55(12):1-38.
<https://doi.org/10.1145/3571730>.

Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *ACM Trans Inf Syst.* 2025;43(2):1-55.
<https://doi.org/10.1145/3703155>.

Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. *Nature.* 2024;630(8017):625-30.
<https://doi.org/10.1038/s41586-024-07421-0>.

Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach Learn.* 2021;110(3):457-506.
<https://doi.org/10.1007/s10994-021-05946-3>.

Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, et al. A review of uncertainty quantification in deep learning: techniques, applications and challenges. *Inf Fusion.* 2021;76:243-97.
<https://doi.org/10.1016/j.inffus.2021.05.008>.

Gawlikowski J, Tassi CRN, Ali M, Lee J, Humt M, Feng J, et al. A survey of uncertainty in deep neural networks. *Artif Intell Rev.* 2023;56(Suppl 1):1513-89.
<https://doi.org/10.1007/s10462-023-10562-9>.

Chaudhari S, Aggarwal P, Murahari V, Rajpurohit T, Kalyan A, Narasimhan K, et al. RLHF deciphered: a critical analysis of reinforcement learning from human feedback for LLMs. *ACM Comput Surv.* 2025;58(2):1-37.
<https://doi.org/10.1145/3743127>.

Gabriel I. Artificial intelligence, values, and alignment. *Minds Mach.* 2020;30(3):411-37.
<https://doi.org/10.1007/s11023-020-09539-2>.

Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A survey on evaluation of large language models. *ACM Trans Intell Syst Technol.* 2024;15(3):1-45.
<https://doi.org/10.1145/3641289>.

Ott S, Barbosa-Silva A, Blagec K, Brauner J, Samwald M. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nat Commun.* 2022;13(1):6793.
<https://doi.org/10.1038/s41467-022-34591-0>.

Shneiderman B. Human-centered artificial intelligence: reliable, safe and trustworthy. *Int J Hum Comput Interact*. 2020;36(6):495-504.

<https://doi.org/10.1080/10447318.2020.1741118>.

Endsley MR. From here to autonomy: lessons learned from human-automation research. *Hum Factors*. 2017;59(1):5-27.

<https://doi.org/10.1177/0018720816681350>.

Raisch S, Krakowski S. Artificial intelligence and management: the automation-augmentation paradox. *Acad Manage Rev*. 2021;46(1):192-210.

<https://doi.org/10.5465/amr.2018.0072>.

Martin K. Ethical implications and accountability of algorithms. *J Bus Ethics*. 2019;160(4):835-50.

<https://doi.org/10.1007/s10551-018-3921-3>.

Morley J, Floridi L, Kinsey L, Elhalal A. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Sci Eng Ethics*. 2020;26(4):2141-68.

<https://doi.org/10.1007/s11948-019-00165-5>.

Lavin A, Gilligan-Lee CM, Visnjic A, Ganju S, Newman D, Ganguly S, et al. Technology readiness levels for machine learning systems. *Nat Commun*. 2022;13(1):6039.

<https://doi.org/10.1038/s41467-022-33128-9>.

Mittelstadt B. Principles alone cannot guarantee ethical AI. *Nat Mach Intell*. 2019;1(11):501-7.

<https://doi.org/10.1038/s42256-019-0114-4>.

Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell*. 2019;1(9):389-99.

<https://doi.org/10.1038/s42256-019-0088-2>.

Song P, Ojo A, Curry E. Trustworthy requirements for foundation models—a comprehensive survey and roadmap. *Eng Appl Artif Intell*. 2026;163(Pt 6):113111.

<https://doi.org/10.1016/j.engappai.2025.113111>.