

ORIGINAL RESEARCH

Open access

Responsible Foundation Models Need More Than Alignment: A Governance and Technical Safeguard Framework for Explainability, Reliability, Human Control, and Institutional Accountability

Daniel Kim¹, Emma Wright^{2*}, Luca Moretti³

Abstract

Foundation models are increasingly embedded in information access, content production, analysis, communication, and decision-support processes. Their broad capabilities have intensified interest in alignment, yet alignment alone cannot establish responsible use. A model may appear responsive to instructions or preferences while remaining opaque, weakly grounded, unreliable under changing conditions, overconfident, difficult to contest, or insufficiently governed. This article proposes a conceptual governance and technical safeguard framework for responsible foundation models. The framework treats foundation models as components of decision infrastructure situated within socio-technical and institutional environments rather than as isolated predictive tools. It integrates eight interdependent areas: model and alignment safeguards; explainability and interpretability; knowledge grounding and provenance; reliability, uncertainty, and hallucination containment; human control and escalation; institutional accountability; auditability and documentation; and lifecycle monitoring and governance updating. Explainability requirements are differentiated from interpretability, transparency, explanation usefulness, and human understanding, while knowledge grounding is separated from factuality, truthfulness, source quality, and provenance. Reliability is addressed through calibration, robustness assessment, boundary testing, uncertainty communication, failure containment, and continuing monitoring. Human oversight is defined through authority, competence, review capacity, escalation pathways, contestability, and recourse rather than nominal human presence. Institutional responsibility is connected to role allocation, evidence trails, documentation, review authority, and consequences. The principal contribution is an integrated framework that clarifies how technical and governance safeguards may support responsible foundation-model assessment without treating alignment, explanation, grounding, testing, oversight, or auditing as independently sufficient. The framework requires empirical validation, domain-specific evaluation, human-centred assessment, institutional governance, and residual-risk management before it can inform consequential operational decisions.

Keywords Explainability, Reliability, Foundation models, Knowledge grounding, Responsible AI, Human control

*Correspondence:

Emma Wright
emma.wright@ed.ac.uk

¹ Department of AI and Governance, Yonsei University, Seoul, South Korea

² Department of Trustworthy AI and Ethics, University of Edinburgh, Edinburgh, United Kingdom

³ Department of Computer Science, University of Milan, Milan, Italy

Introduction

Foundation models have expanded generative artificial intelligence from specialised prediction toward reusable systems capable of producing, transforming, summarising, and organising multiple forms of information. Their adaptability allows a single underlying model to mediate diverse tasks, interfaces, and organisational processes, but this breadth also increases the number of contexts in which errors, omissions, misleading outputs, or inappropriate automation may become consequential. Responsible foundation-model use therefore concerns not only model behaviour but also how generated information enters decisions, how uncertainty is interpreted, and how responsibility is distributed around the system [1].

The infrastructural significance of foundation models becomes especially visible when models support multiple tasks, modalities, and professional functions. Generalist architectures may combine text, images, structured information, and domain knowledge within shared workflows, allowing generated outputs to influence interpretation, prioritisation, communication, and subsequent action. Such capacity does not itself demonstrate that outputs are reliable, suitable, or responsibly governed. Instead, broader functionality enlarges the range of downstream dependencies and requires evaluation at the level of the integrated decision environment rather than through capability demonstrations alone [2].

Alignment is necessary within this environment because model behaviour must be shaped toward specified objectives, constraints, and expectations. However, responsible use cannot be inferred merely from the presence of behavioural objectives or ethical principles. High-level commitments remain indeterminate unless institutions translate them into operational responsibilities, evidence requirements, review mechanisms, and enforceable practices. A system may appear aligned with selected preferences while still obscuring uncertainty, drawing on unreliable information, generating unsupported claims, or operating without meaningful accountability [3].

Trustworthiness is consequently better understood as a multidimensional socio-technical condition than as a singular model attribute. It involves technical properties, human interaction, organisational processes, institutional authority, and continuing evaluation across the system lifecycle. This article develops an original governance and technical safeguard framework that connects alignment limits with explainability, knowledge grounding, reliability, uncertainty communication, failure containment, human control, accountability, auditability, and lifecycle monitoring. The framework structures responsible-use questions and decision boundaries but does not independently establish safety, ethical sufficiency, legal compliance, certification, or deployment readiness [4].

Foundation Models as Decision Infrastructure

Foundation models constitute a distinctive artificial-intelligence paradigm because broadly trained models can be adapted, prompted, or integrated into many downstream applications. Large language models primarily process and generate linguistic representations, while large multimodal models connect text with images or other data modalities. Generative AI describes the broader capacity to produce new content, and general-purpose AI emphasises applicability across tasks and sectors. These categories overlap but are not interchangeable. Their shared relevance lies in the ability of reusable models to become embedded within multiple technical and organisational systems [5].

Once integrated into workflows, a foundation model may influence how information is searched, summarised, classified, interpreted, or presented to decision makers. It can therefore shape the architecture of decision-making even when it has no formal authority to make a final decision. Organisational arrangements determine whether a model automates a task, advises a professional, filters available options, or structures the information on which others rely. Responsibility must accordingly be mapped across model providers, system integrators, organisational managers, domain professionals, reviewers, and other actors who determine how outputs are used [6].

Treating foundation models as decision infrastructure does not require assuming that humans will be displaced. Human judgement may complement computational capacity by interpreting context, recognising exceptions, considering tacit knowledge, resolving value conflicts, and assuming responsibility for consequential choices. Nevertheless, complementarity is not automatic. It depends on whether humans receive sufficient information, time, authority, expertise,

and alternative options to evaluate model-generated material. Human participation that consists only of approving fluent output may reproduce automation dependence rather than create substantive human–AI collaboration [7].

The distinction between automation and augmentation is therefore organisationally unstable rather than absolute. A tool introduced to support human work may gradually define available options, standardise reasoning, increase workload expectations, or reduce opportunities for independent judgement. Conversely, an automated process may still require extensive human supervision, correction, and institutional support. Model scale, linguistic fluency, apparent generality, or convenience must not be treated as evidence of responsibility. The framework instead requires explicit analysis of the model's role, decision influence, dependencies, risk transfer, affected stakeholders, and institutional embedding [8].

Explainability and Knowledge-Grounding Requirements

Explainability is an umbrella concept covering multiple ways of making model behaviour, outputs, or decision contributions more intelligible. Interpretability concerns the degree to which relevant mechanisms or relationships can be understood, whereas transparency concerns the availability of material information about the model, data, processes, limitations, and organisational use. Post-hoc explanations may provide feature attributions, examples, summaries, rules, or counterfactual accounts, while mechanistic interpretability may examine internal computational structures where technically supported. These approaches have different evidential strengths and cannot be treated as equivalent safeguards [9].

Explanation quality must be assessed in relation to purpose and recipient. An explanation intended for a developer investigating model behaviour differs from one intended for a professional reviewer, end user, auditor, or affected stakeholder. Useful explanations should be relevant, cognitively accessible, appropriately selective, and capable of supporting the recipient's task without concealing material limitations. They should also communicate explanation uncertainty, because a coherent explanation may still be incomplete, approximate, unstable, or weakly faithful to the model process. Human understanding must therefore be evaluated rather than presumed from the mere availability of explanatory material [10].

In high-impact settings, post-hoc explanations may be insufficient when they offer plausible narratives around a model whose relevant behaviour remains difficult to inspect. Interpretability can sometimes provide stronger grounds for review, particularly when decision logic must be directly examined, but interpretable representations are not automatically accurate, fair, or contextually appropriate. The framework consequently imposes a no-explainability-sufficiency boundary: neither interpretability nor post-hoc explanation independently demonstrates reliability, legitimate decision-making, safety, or accountability [11].

Knowledge grounding addresses a different problem by connecting generated claims to identifiable external information. Retrieval augmentation, citation grounding, and controlled knowledge sources may help constrain generation and improve verifiability, but their effectiveness depends on retrieval quality, source authority, currency, completeness, and claim-to-source correspondence. Groundedness indicates whether a claim is supported by supplied context, whereas factuality and truthfulness require broader assessment of whether the claim and its sources are accurate. Knowledge provenance and data provenance must document origins, versions, transformations, and limitations. Grounding therefore reduces some unsupported-generation risks without guaranteeing that retrieved information is true or that hallucination has been eliminated [12]. **Table 1** summarises explainability and knowledge-grounding requirements for responsible foundation models, including explanation quality, grounding, provenance, factuality, uncertainty, validation, and safeguard boundaries.

Table 1. Explainability and Knowledge-Grounding Requirements for Responsible Foundation Models: Interpretability, Transparency, Explanation Quality, Explanation Usefulness, Explanation Uncertainty, Human Understanding, Knowledge Grounding, Factuality, Groundedness, Truthfulness, Knowledge Provenance, Data Provenance, Retrieval Augmentation, Source Validation, and Sufficiency Boundaries

Requirement	Core responsible-use question	Required evidence or design condition	Safeguard function	Potential output	Failure or overclaiming risk	Validation or governance need
Explainability	Can relevant actors obtain an account suited to their decision task?	Defined audience, purpose, timing, scope, and decision relevance	Supports review and challenge	Context-specific explanation	Treating any explanation as proof of correctness	Human-centred and domain evaluation
Interpretability	Can relevant behaviour be examined without relying exclusively on a post-hoc narrative?	Interpretable representation or justified model-choice rationale	Supports direct examination	Comprehensible model evidence	Equating simplicity with safety	Technical validation
Transparency	Is material information about purpose, data, limitations, and use available?	Purposeful, accessible, and appropriately scoped disclosure	Supports scrutiny	System or process disclosure	Transparency theatre or information overload	Disclosure governance
Explanation quality	Is the explanation faithful, relevant, stable, and sufficiently complete?	Fidelity, sensitivity, consistency, and relevance assessment	Reduces misleading explanations	Quality-assessed explanation	Plausible but unfaithful reasoning	Method-specific validation
Explanation usefulness	Does the explanation support the recipient's actual task?	Task-fit, actionability, and user-comprehension evidence	Supports informed use	Task-relevant explanation	Readable but decision-irrelevant output	Human-centred assessment
Explanation uncertainty	Are explanatory limitations communicated?	Scope conditions, uncertainty indicators, and qualification	Prevents false certainty	Qualified explanation	Excessive precision or unjustified confidence	Communication evaluation
Mechanistic and post-hoc interpretation	Is the selected method appropriate to the risk and question?	Validated probes, attributions, examples, rules, or counterfactuals	Supports investigation at different levels	Mechanistic or local explanatory evidence	Inferring global understanding from partial evidence	Reproducibility and fidelity review

		with declared limitations				
Human understanding and cognitive accessibility	Can intended recipients correctly understand and use the information?	Comprehension testing, appropriate complexity, accessible language, and alternative formats	Reduces misunderstanding and exclusion	Demonstrated recipient understanding	Assuming exposure equals comprehension	Accessibility and stakeholder evaluation
Knowledge grounding and grounded generation	Are generated claims connected to identifiable supporting material?	Source-linked generation and claim–evidence alignment	Constrains unsupported generation	Evidence-linked output	Citation decoration or irrelevant retrieval	Claim-level grounding evaluation
Factuality, groundedness, and truthfulness	Are claims accurate, supported by context, and appropriately qualified?	Fact checking, entailment assessment, and evidence hierarchy	Identifies epistemic failure	Factuality and groundedness record	Treating context support as proof that a source is true	Independent domain validation
Knowledge and data provenance	Can information and data origins, versions, and transformations be traced?	Lineage records, permissions, timestamps, and processing documentation	Supports verification and accountability	Provenance trail	Incomplete or unverifiable lineage	Data and knowledge stewardship
Retrieval augmentation and citation grounding	Does retrieval obtain suitable evidence and do citations support associated claims?	Validated corpus, retrieval testing, access controls, source checking, and citation audit	Supplies external knowledge at generation time	Retrieved evidence set with citations	Obsolete, manipulated, fabricated, or mismatched sources	Continuing corpus and retrieval monitoring
Knowledge-source validation	Are sources authoritative, relevant, current, and intact?	Source-selection criteria, integrity checks, and ownership	Prevents unreliable evidence from appearing authoritative	Approved source record	Authoritative-looking misinformation	Knowledge-governance review
No explainability- or grounding-sufficiency boundary	What remains unresolved after explanations or sources are provided?	Explicit residual-risk statement separating explanation and grounding	Prevents safeguard overclaiming	Decision-boundary statement	Explanation-washing or grounding-washing	Cross-layer governance review

		from safety, truth, legitimacy, and compliance				
--	--	---	--	--	--	--

Reliability, Alignment, and Failure Containment

Alignment concerns the relationship between model behaviour and specified goals, instructions, preferences, constraints, or values. Human alignment and value alignment additionally require questions about whose interests are represented, how disagreements are resolved, and whether the procedures used to shape behaviour are legitimate. Within the proposed framework, instruction tuning, preference learning, reinforcement learning from human feedback, and safety tuning are treated as behaviour-shaping mechanisms where present, not as independent evidence of responsibility. Alignment evaluation must therefore document objectives, preference sources, affected groups, unresolved conflicts, and conditions outside the available evidence [13].

Reliability concerns whether a system performs dependably for its defined role and operating conditions. Confidence expressions generated in natural language should not be assumed to represent calibrated probabilities. Calibration instead requires evidence that estimated confidence corresponds meaningfully to observed correctness within relevant contexts. Uncertainty estimation can support review and escalation, but only when the method has been validated for the task and population. Uncertainty communication must also be tested with intended users, because even technically informative signals may induce overreliance, avoidance, or misunderstanding [14].

Hallucination and fabrication are central reliability concerns because a model may generate unsupported claims, invented entities, false citations, or misleadingly specific information. Semantic uncertainty can help identify some cases in which multiple plausible generations reflect unstable meaning, but detection methods remain partial and context dependent. The framework therefore combines hallucination-risk assessment with grounding, factuality checking, citation verification, abstention where appropriate, and escalation for consequential claims. These mechanisms support containment; they do not establish that hallucination has been eliminated [15].

Robustness assessment must examine how performance changes under perturbations, adversarial inputs, distribution shifts, altered prompts, missing information, and unfamiliar contexts. Stress testing and red-teaming may reveal boundary failures or misuse pathways where supported by the system context, but one test set or red-team exercise cannot prove general safety. Testing should cover data, model behaviour, interfaces, system integration, and downstream decision effects. Failure-containment controls should then restrict the ability of unreliable output to trigger consequential action, while escalation thresholds should direct uncertain or high-impact cases to qualified review [16].

Reliability and failure containment must continue after initial evaluation. Model updates, changing data, revised knowledge sources, new user practices, organisational adaptation, and environmental change can invalidate earlier evidence. Lifecycle assurance therefore requires version control, monitoring, incident detection, evidence preservation, change assessment, and governance review. Incident-response planning may define containment, investigation, correction, communication, and framework updating where appropriate. The proposed framework maintains three explicit boundaries: alignment does not establish responsibility, reliability evaluation does not guarantee safety outside validated conditions, and hallucination controls do not prove the impossibility of future unsupported generation [17].

Human Control and Institutional Accountability

Human control requires more than placing a person somewhere within an automated workflow. Oversight becomes meaningful only when the responsible person possesses sufficient information, competence, time, authority, and practical capacity to affect the outcome. Human-in-the-loop arrangements require affirmative review before action, whereas

human-on-the-loop arrangements permit supervised automation with intervention capacity. The suitability of either arrangement depends on the decision context, system speed, consequences, and reviewer capabilities. Nominal human presence cannot compensate for inaccessible reasoning, excessive workload, or absent override authority [18].

Human-centred design can combine substantial automation with substantial human control when system functions are deliberately organised around human agency. Reviewers need understandable outputs, uncertainty information, access to relevant evidence, and the ability to reject, revise, or escalate model recommendations. Override capacity may be necessary where model-generated content can initiate consequential processes, although an override mechanism is useful only when it is accessible, tested, and assigned to an authorised actor. Human control should therefore be designed as an operational capability rather than a symbolic assurance claim [19].

Institutional accountability extends beyond identifying an individual who interacted with the model. An accountable institution must be able to explain why the system was used, which safeguards were applied, who authorised material decisions, how challenges are reviewed, and what consequences follow from failures. Answerability requires evidence and review mechanisms, while enforceability requires authority to correct practices, restrict use, or impose consequences. Accountability is weakened when organisations attribute adverse outcomes to technical complexity, autonomous behaviour, or external model providers without examining their own deployment decisions [20].

Organisational accountability requires explicit responsibility mapping across development, procurement, integration, data stewardship, operational use, human review, incident handling, and governance oversight. Role allocation should distinguish technical responsibility from decision authority, domain judgement, records management, and institutional approval. Model developers may document capabilities and limitations, but deployers remain responsible for local integration and use. System integrators and operational managers likewise retain responsibilities for interface design, escalation rules, downstream dependencies, and the *organisational conditions under which model output becomes influential* [21].

Auditability, traceability, and documentation support accountability but do not automatically create it. Model documentation should identify versions, capabilities, limitations, and evaluation scope; data documentation should preserve provenance and processing information; and decision documentation should record consequential uses, evidence, reviewers, and dispositions. Evidence trails and protected audit trails can assist reconstruction, investigation, contestability, and recourse. Nevertheless, transparency without interpretive capacity, institutional authority, or corrective consequences may merely reveal information without enabling meaningful accountability. The framework therefore maintains a no-human-control-sufficiency boundary and a no-accountability-guarantee boundary [22].

Technical and Governance Safeguards

Model-level safeguards address behaviour, access, evaluation, versioning, and alignment limitations, while data-level safeguards govern quality, provenance, permissions, and change. Foundation-model responsibility also depends on knowledge-source management because retrieval systems can propagate obsolete, incomplete, manipulated, or contextually unsuitable material. Data and knowledge governance should assign stewardship roles, document lineage, establish validation processes, and detect changes that could undermine grounding or reliability. These arrangements connect technical controls to accountable organisational functions [23].

Explanation, grounding, reliability, uncertainty, and hallucination-containment safeguards address different failure pathways. Explanation safeguards support interpretation and review; grounding safeguards connect claims to sources; reliability safeguards evaluate consistency and boundary behaviour; uncertainty safeguards support calibrated reliance; and containment safeguards restrict the consequences of unsupported output. Practical responsible-AI methods may assist these functions, but available tools differ in scope, evidence, maturity, and suitability. A tool should therefore be selected and evaluated according to a defined risk question rather than treated as a generic implementation of ethical principles [24].

Auditability safeguards require specified claims, evaluative criteria, accessible evidence, suitable independence, and acknowledgement of scope limitations. An audit may assess whether documented procedures were followed, whether selected risks were examined, or whether evidence supports particular governance claims. It cannot independently determine that a system is ethically sufficient, universally acceptable, legally compliant, or operationally safe. Audit findings should therefore feed into broader institutional deliberation, risk management, and corrective action rather than function as certification by default [25].

Responsible foundation-model safeguards must combine technical controls, human-control mechanisms, and institutional accountability rather than relying on alignment alone. Technical measures can shape model behaviour and detect some errors, but organisations determine purposes, thresholds, authority, escalation, documentation, and acceptable residual risk. Layered governance connects model, data, knowledge, human, organisational, and societal functions while recognising that no individual layer can substitute for the others. The architecture must also permit information and failure signals to move upward and corrective decisions to move back toward technical and operational layers [26].

Figure 1 illustrates the layered safeguard architecture needed for responsible foundation models, connecting alignment limits, explainability, grounding, reliability, human control, accountability, monitoring, and governance review.

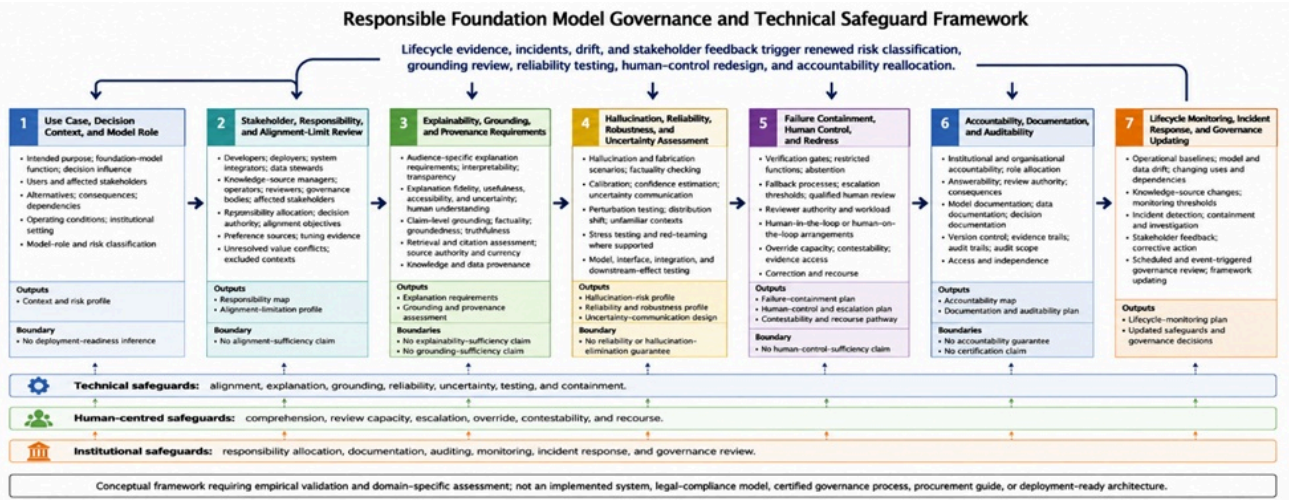


Figure 1. Layered Safeguard Architecture for Responsible Foundation Models

Alt-text description:

A layered safeguard architecture showing alignment limits, explainability, knowledge grounding, reliability, hallucination containment, uncertainty communication, human control, institutional accountability, auditability, monitoring, and governance updating.

The safeguard architecture must remain dynamic because trustworthy-AI properties can interact, conflict, or deteriorate as systems and environments change. Strong grounding may improve evidential traceability without resolving value conflicts; extensive documentation may support auditing without ensuring effective recourse; and human review may detect some errors while introducing inconsistency or automation bias. Lifecycle-monitoring and governance-review safeguards should therefore assess interactions among controls, emerging failure modes, model and data drift, stakeholder feedback, and incidents. The resulting evidence supports continuing risk management, not a declaration that the system has become universally trustworthy [27].

Table 2 maps reliability, alignment, failure containment, human control, institutional accountability, and lifecycle governance safeguards for responsible foundation model use.

Table 2. Technical and Governance Safeguards for Responsible Foundation Models: Alignment Limits, Reliability, Robustness, Uncertainty, Hallucination Containment, Failure Containment, Human Control, Escalation, Contestability, Institutional Accountability, Auditability, Documentation, Lifecycle Monitoring, Incident Response, and Safeguard Boundaries

Safeguard layer or requirement	Core safeguard question	Technical or governance function	Required evidence or process	Failure risk addressed	Residual risk	Monitoring or review need	Fr
Model and alignment safeguards	Does behaviour remain within defined objectives and limitations?	Technical and governance	Version-specific evaluation, objective documentation, access controls, and limitation review	Harmful, unstable, or misaligned behaviour	Misspecified goals and unseen capabilities	Repeated behavioural and adversarial evaluation	Mo
Data safeguards	Are data suitable, traceable, and legitimately governed?	Technical and governance	Provenance, quality review, permissions, stewardship, and version control	Bias, contamination, poor quality, and invalid use	Hidden historical or contextual bias	Data-quality and drift review	D
Knowledge-grounding safeguards	Are generated claims connected to suitable evidence?	Technical and governance	Source validation, retrieval testing, provenance, and claim–source assessment	Unsupported or obsolete generation	False or incomplete sources	Source and retrieval monitoring	Ki
Explanation safeguards	Can relevant actors understand and assess system contributions?	Technical and human-centred	Explanation requirements, fidelity assessment, accessibility, and user testing	Opacity and misleading explanation	Plausible but unfaithful accounts	Audience-specific evaluation	Ex
Reliability and robustness safeguards	Does performance remain dependable under relevant conditions?	Technical	Validation, calibration, perturbation testing, stress testing, and boundary analysis	Inconsistent or fragile behaviour	Unseen conditions and interactions	Operational performance review	R

Uncertainty safeguards	Can uncertainty be estimated and communicated usefully?	Technical and human-centred	Calibration evidence, uncertainty method validation, and comprehension testing	Overconfidence and uncalibrated reliance	Undetected uncertainty under shift	Calibration and interface review	Review of uncertainty
Hallucination and fabrication containment	Can unsupported content be detected, verified, restricted, or escalated?	Technical and operational	Grounding, factuality review, citation checking, abstention, and escalation	Fabricated or misleading information	Undetected hallucination	Sampling and incident analysis	Contestability and recourse
Failure-containment and incident response	Can failures be prevented from propagating and addressed when detected?	Technical, operational, and governance	Isolation, validation gates, fallback processes, reporting, investigation, and remediation	Consequential failure propagation	Common-mode control failure	Exercises and post-incident review	Contestability and recourse
Human-control safeguards	Do qualified humans possess effective authority and intervention capacity?	Human-centred and governance	Review protocols, escalation thresholds, override mechanisms, competence, and workload controls	Unreviewed consequential automation	Automation bias and reviewer failure	Control-effectiveness evaluation	Incident response and containment
Contestability and recourse	Can affected stakeholders challenge and obtain correction?	Governance	Notice, reasons, evidence access, review, and remedial authority	Unchallengeable adverse outcomes	Unequal access or ineffective remedies	Appeal and outcome review	Review of uncertainty
Institutional accountability	Who must answer for decisions and consequences?	Governance	Responsibility mapping, reporting authority, review mechanisms, and consequences	Responsibility displacement	Weak enforcement	Governance-effectiveness review	Incident response and containment

Auditability and traceability	Can relevant claims, events, and changes be independently examined?	Technical and governance	Criteria, logs, evidence access, versioning, and independence	Unverifiable claims and lost evidence	Audit gaming or incomplete scope	Periodic independent examination	A
Documentation safeguards	Are model, data, decision, and governance records maintained?	Governance	Versioned documentation and controlled evidence repositories	Institutional memory loss and opaque change	Incomplete or outdated records	Documentation review	E
Lifecycle monitoring	Are drift, changing use, performance, and emerging risks detected?	Technical and governance	Baselines, indicators, ownership, thresholds, and review triggers	Degradation and function creep	Slow or unobservable harm	Continuing monitoring	Life
Governance review	Are assumptions, purposes, risks, and safeguards periodically reconsidered?	Governance	Scheduled and event-triggered review with stakeholder input	Stale governance decisions	Weak institutional authority	Independent and participatory review	Upd
No single-safeguard sufficiency boundary	Is any control being treated as proof of responsibility, safety, or compliance?	Cross-cutting	Integrated evidence review and residual-risk register	Alignment-, explanation-, grounding-, audit-, or oversight-washing	Unanticipated interactions among safeguards	Cross-layer governance assessment	b

Proposed Responsible Foundation Model Framework

The proposed framework begins with use-case and decision-context definition. Institutions should identify the intended purpose, decision influence, affected stakeholders, alternatives, consequences, and conditions under which the foundation model will operate. Foundation-model role and risk classification then distinguish whether the model generates information, recommends options, filters evidence, supports professional judgement, or initiates action. This contextual stage translates broad responsible-AI principles into questions about actual institutional use rather than assuming that general principles automatically determine appropriate safeguards [28].

The second stage maps stakeholders and responsibilities before examining alignment limitations. Developers, deployers, integrators, data stewards, knowledge-source managers, operators, reviewers, governance bodies, end users, and affected stakeholders may hold different forms of authority and knowledge. Alignment evidence must identify behavioural objectives, preference sources, unresolved value conflicts, and excluded contexts. Broad ethical guidance commonly endorses transparency, safety, responsibility, and human autonomy, but variation in interpretation demonstrates the need to specify how each principle is operationalised within the particular use context [29].

The third stage defines explainability requirements, evaluates knowledge grounding and provenance, and assesses hallucination and factuality risks. Explanation purposes should be assigned to specific audiences, while grounding should be evaluated through source quality, retrieval performance, claim-level support, and provenance. Reliability and robustness assessment should then examine calibrated confidence, uncertainty, perturbations, distribution shift, and failure conditions. Ethical guidance that lacks technical implementation detail may leave these requirements too abstract to support review, making explicit evidence conditions and decision boundaries necessary [30].

The fourth stage designs failure containment, human control, escalation, contestability, recourse, institutional accountability, documentation, and auditability. Technical methods should be linked to defined trustworthiness objectives, while human and organisational safeguards should determine how evidence is interpreted and acted upon. High-risk or uncertain cases may require restricted system functions, verification gates, qualified human review, escalation, or fallback processes. These components remain conceptual until tested in a specific domain and should not be interpreted as universal technical configurations or operational instructions [31].

The final stage establishes lifecycle monitoring, model-drift and data-drift review, incident detection and response where supported, governance review, and framework updating. Monitoring evidence should feed back into risk classification, grounding arrangements, reliability assessment, human-control design, and accountability allocation. The framework therefore operates as a revisable governance cycle rather than a one-time checklist. It connects technical practices with organisational mechanisms while preserving explicit boundaries against alignment sufficiency, explanation sufficiency, grounding sufficiency, reliability guarantees, human-control sufficiency, accountability guarantees, compliance claims, and deployment-readiness claims [32].

Table 3 presents the proposed responsible foundation model framework linking decision context, alignment limits, explainability, grounding, reliability, human control, accountability, safeguards, monitoring, and governance updating.

Table 3. Proposed Responsible Foundation Model Governance and Technical Safeguard Framework: Use-Case Definition, Model Role, Risk Context, Responsibility Mapping, Alignment-Limit Review, Explainability Requirements, Knowledge-Grounding Review, Hallucination Risk, Reliability Assessment, Uncertainty Communication, Failure Containment, Human Control, Contestability, Accountability, Auditability, Lifecycle Monitoring, and Deployment Boundaries

Framework stage	Core purpose	Required input	Safeguard function	Potential output	Validation or governance need	Failure risk
Use-case and decision-context definition	Establish purpose, stakes, and affected actors	Intended use, decisions, users, consequences, and alternatives	Prevents vague or expanding use	Context statement	Domain and stakeholder review	Function creep
Model role and risk classification	Determine the model's decision influence	Architecture, integrations, authority, and impact	Links role to safeguard intensity	Role and risk classification	Institutional review	Underclassification
Stakeholder and responsibility mapping	Assign roles, authority, and affected-party recognition	Stakeholder analysis and organisational structure	Prevents responsibility gaps	Responsibility map	Governance approval	Excluded stakeholders or unclear ownership

Alignment-limit review	Identify what behaviour-shaping evidence covers	Objectives, preference sources, tuning records, and evaluation	Prevents alignment overclaiming	Alignment limitation profile	Independent challenge review	Alignment-washing
Explainability requirement definition	Specify who requires what form of explanation	Audience, task, timing, and contestability needs	Prevents generic explanation	Explanation requirements	Human-centred validation	Unusable or unfaithful explanation
Knowledge-grounding and provenance review	Assess external evidence and lineage	Source corpus, retrieval process, provenance, and update controls	Supports evidential traceability	Grounding assessment	Knowledge-steward review	False or obsolete grounding
Hallucination and factuality assessment	Identify unsupported-generation risks	Claim types, evidence standards, and failure scenarios	Prioritises verification and containment	Hallucination-risk profile	Domain validation	Undetected fabrication
Reliability and robustness assessment	Characterise dependable operation	Validation evidence, perturbations, scenarios, and metrics	Tests performance and boundaries	Reliability profile	Independent technical review	Benchmark overgeneralisation
Uncertainty communication design	Support calibrated human reliance	Calibration evidence, uncertainty method, and interface options	Makes uncertainty actionable	Communication design	Comprehension testing	Misleading confidence
Failure-containment design	Restrict the consequences of erroneous output	Failure modes, dependencies, action pathways, and fallback options	Limits failure propagation	Containment design	Scenario testing	Simultaneous safeguard failure
Human-control and escalation design	Preserve review and intervention capacity	Decision rights, expertise, thresholds, workload, and response time	Routes consequential or uncertain cases	Control and escalation plan	Operational evaluation	Rubber-stamping or unavailable reviewers

Contestability and recourse design	Enable challenge, correction, and remedy	Notice, reasons, evidence access, and remedial authority	Protects affected stakeholders	Challenge and recourse pathway	Institutional governance	Formal review without meaningful remedy
Institutional accountability mapping	Establish answerability and consequences	Governance mandate, ownership, and review authority	Prevents responsibility displacement	Accountability map	Senior institutional approval	Diffuse or symbolic accountability
Documentation and auditability design	Preserve reviewable evidence	Documentation schema, logs, versions, and access controls	Enables traceability and examination	Evidence plan	Records and assurance review	Missing or mutable evidence
Lifecycle monitoring and drift review	Detect changing performance, context, or validity	Baselines, indicators, thresholds, operational data, and owners	Extends safeguards beyond initial evaluation	Monitoring and drift plan	Operational governance	Undetected degradation
Incident response and governance updating	Contain failures and revise the framework	Reports, investigation evidence, stakeholder input, and corrective findings	Supports learning and corrective governance	Updated safeguards and decisions	Independent or multistakeholder review	Delayed response or governance inertia
Compliance and deployment boundaries	Separate framework completion from legal or operational authorisation	Domain evidence, residual-risk decision, legal analysis, and institutional authority	Prevents checklist-based approval	Explicit limitation statement	Separate qualified review	Premature use or unsupported compliance claims

Figure 2 presents the proposed responsible foundation model framework, integrating decision-context definition, alignment-limit review, explainability, grounding, reliability, failure containment, human control, accountability, lifecycle monitoring, and governance updating.

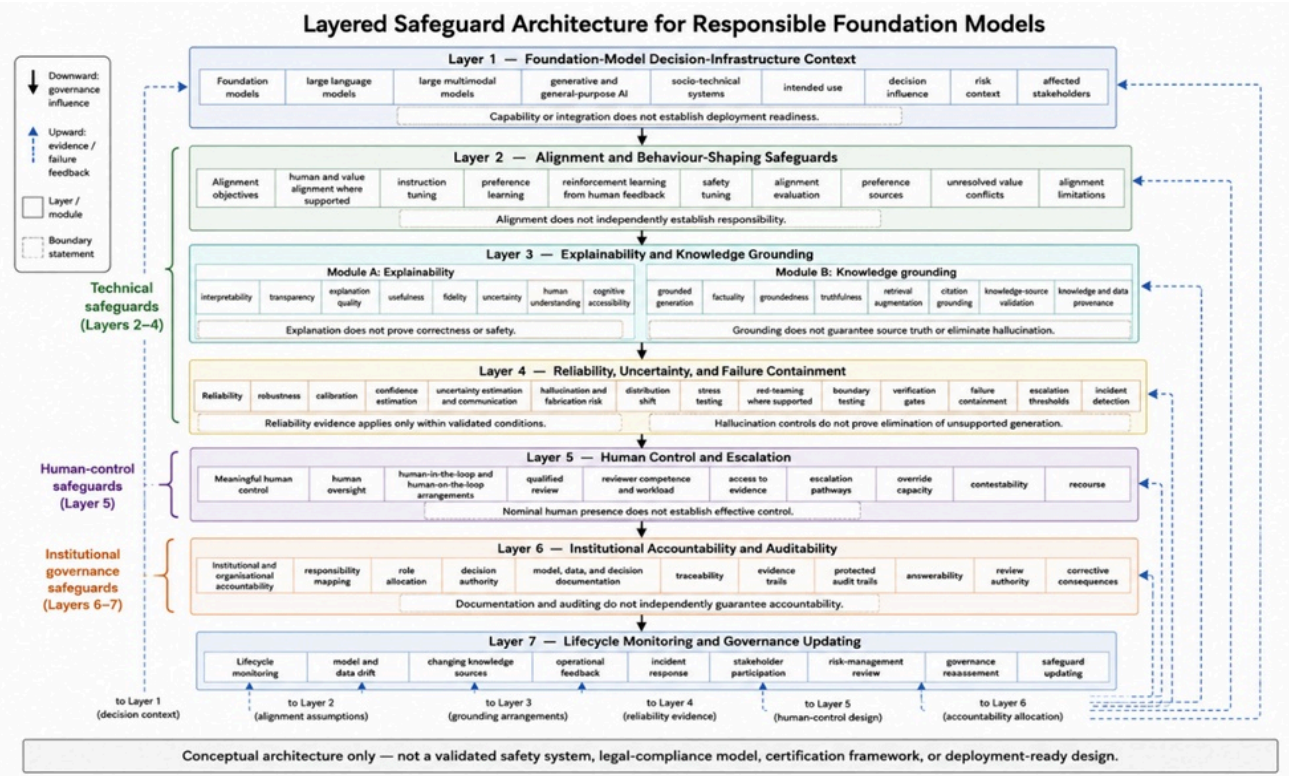


Figure 2. Responsible Foundation Model Governance and Technical Safeguard Framework

Alt-text description:

A conceptual responsible foundation model framework showing use-case definition, alignment limits, explainability, knowledge grounding, hallucination risk, reliability, uncertainty, failure containment, human control, accountability, documentation, auditability, monitoring, incident response, and governance updating.

Practical, Policy, and Research Implications

For developers, deployers, and system integrators, the framework clarifies that model evaluation should be connected to the intended decision role and surrounding organisational controls. Developers can provide version-specific evidence about capabilities, limitations, alignment methods, uncertainty, and known failure conditions, while deployers and integrators must evaluate local data, interfaces, knowledge sources, dependencies, review processes, and escalation arrangements. These implications do not prescribe particular models, platforms, vendors, architectures, procurement decisions, or deployment configurations.

Data stewards and knowledge-source managers require stronger methods for provenance, source validation, change management, retrieval evaluation, and claim-level grounding assessment. Researchers should distinguish factuality from groundedness and examine how source error, retrieval failure, citation mismatch, and model fabrication interact. Explainability research should likewise integrate technical fidelity with usefulness, cognitive accessibility, explanation uncertainty, and stakeholder-specific evaluation. Evaluation should focus on whether explanations and grounding support appropriate review rather than merely producing plausible supporting material.

Human reviewers, decision-support operators, institutional governance bodies, and auditors require evidence about actual control effectiveness. Research should assess whether reviewers understand model limitations, respond appropriately to uncertainty, possess sufficient authority, and can use escalation and override pathways under realistic

workload conditions. Accountability research should examine whether responsibility maps produce answerability, correction, learning, and consequences. Auditing research should investigate evidence quality, independence, scope selection, and the risks of reducing governance to procedural conformity.

Policy and governance research should compare how technical and organisational safeguards function across sectors, institutions, and decision contexts. Lifecycle studies are needed to examine drift, changing human dependence, emerging uses, incident patterns, safeguard interactions, and residual-risk decisions. The proposed framework may provide common analytical dimensions for such research, but it does not constitute legal advice, compliance guidance, procurement advice, audit instruction, model-selection advice, platform-selection advice, policy prescription, safety certification, or deployment instruction.

Conclusion

Responsible foundation models require more than behavioural alignment because model outputs become consequential through explanation practices, knowledge sources, reliability conditions, human decisions, organisational structures, and institutional authority. The proposed framework integrates explainability, interpretability, grounding, provenance, factuality, uncertainty communication, robustness, hallucination containment, failure containment, human control, escalation, contestability, accountability, auditability, documentation, monitoring, incident response, and governance updating. It rejects the use of alignment claims, fluent output, benchmark performance, model scale, transparency, grounding, testing, oversight, or auditing as independent evidence of responsibility. The framework instead supports structured, context-specific inquiry into safeguard evidence and residual risk. It remains conceptual and requires empirical validation, domain-specific assessment, stakeholder evaluation, institutional governance, and continuing lifecycle review.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 14 Oct 2025

Revised: 31 Oct 2025

Accepted: 08 Dec 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

- 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>.
- 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>.
- Mittelstadt BD. Principles alone cannot guarantee ethical AI. *Nat Mach Intell*. 2019;1(11):501-7.
<https://doi.org/10.1038/s42256-019-0114-4>.
- 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>.
- Schneider J, Meske C, Kuss P. Foundation models: A new paradigm for artificial intelligence. *Bus Inf Syst Eng*. 2024;66(2):221-31.
<https://doi.org/10.1007/s12599-024-00851-0>.
- Shrestha YR, Ben-Menahem SM, von Krogh G. Organizational decision-making structures in the age of artificial intelligence. *Calif Manage Rev*. 2019;61(4):66-83.
<https://doi.org/10.1177/0008125619862257>.
- Jarrahi MH. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Bus Horiz*. 2018;61(4):577-86.
<https://doi.org/10.1016/j.bushor.2018.03.007>.
- 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>.
- Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot 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>.
- 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>.
- Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-15.
<https://doi.org/10.1038/s42256-019-0048-x>.
- Klesel M, Wittmann HF. Retrieval-augmented generation (RAG). *Bus Inf Syst Eng*. 2025;67(4):551-61.
<https://doi.org/10.1007/s12599-025-00945-3>.
- Gabriel I. Artificial intelligence, values, and alignment. *Minds Mach*. 2020;30(3):411-37.
<https://doi.org/10.1007/s11023-020-09539-2>.
- Jiang Z, Araki J, Ding H, Neubig G. How can we know when language models know? On the calibration of language models for question answering. *Trans Assoc Comput Linguist*. 2021;9:962-77.
https://doi.org/10.1162/tac1_a_00407.

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>.

Zhang JM, Harman M, Ma L, Liu Y. Machine learning testing: Survey, landscapes and horizons. *IEEE Trans Softw Eng*. 2022;48(1):1-36.
<https://doi.org/10.1109/TSE.2019.2962027>.

Ashmore R, Calinescu R, Paterson C. Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ACM Comput Surv*. 2021;54(5):1-35.
<https://doi.org/10.1145/3453444>.

Christen M, Burri T, Kandul S, Vörös P. Who is controlling whom? Reframing meaningful human control of AI systems in security. *Ethics Inf Technol*. 2023;25(1):10.
<https://doi.org/10.1007/s10676-023-09686-x>.

Shneiderman B. Human-centered artificial intelligence: Reliable, safe & trustworthy. *Int J Hum Comput Interact*. 2020;36(6):495-504.
<https://doi.org/10.1080/10447318.2020.1741118>.

Busuioc M. Accountable artificial intelligence: Holding algorithms to account. *Public Adm Rev*. 2021;81(5):825-36.
<https://doi.org/10.1111/puar.13293>.

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>.

Ananny M, Crawford K. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media Soc*. 2018;20(3):973-89.
<https://doi.org/10.1177/1461444816676645>.

Janssen M, Brous P, Estevez E, Barbosa LS, Janowski T. Data governance: Organizing data for trustworthy artificial intelligence. *Gov Inf Q*. 2020;37(3):101493.
<https://doi.org/10.1016/j.giq.2020.101493>.

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>.

Mökander J, Morley J, Taddeo M, Floridi L. Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations. *Sci Eng Ethics*. 2021;27(4):44.
<https://doi.org/10.1007/s11948-021-00319-4>.

Gasser U, Almeida VAF. A layered model for AI governance. *IEEE Internet Comput*. 2017;21(6):58-62.
<https://doi.org/10.1109/MIC.2017.4180835>.

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>.

Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds Mach*. 2018;28(4):689-707.
<https://doi.org/10.1007/s11023-018-9482-5>.

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>.

Hagendorff T. The ethics of AI ethics: An evaluation of guidelines. *Minds Mach.* 2020;30(1):99-120.
<https://doi.org/10.1007/s11023-020-09517-8>.

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>.

Papagiannidis E, Mikalef P, Conboy K. Responsible artificial intelligence governance: A review and research framework. *J Strateg Inf Syst.* 2025;34(2):101885.
<https://doi.org/10.1016/j.jsis.2024.101885>.