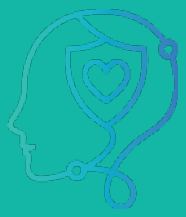

RECOMMENDED PRACTICES

Safer Agentic AI Foundations

A comprehensive framework for developing, deploying, and governing artificial intelligence systems with autonomous agency — establishing safety foundations through drivers and inhibitors.

) | August 2026

unity of Practice

9 Drivers | 7 Inhibitors | 801 Requirements

www.SaferAgenticAI.org

Table of Contents

1. Overview

> Welcome to the Safer Agentic AI Foundations framework — a collaborative effort to establish safety standards for artificial intelligence systems with autonomous agency.

Dear Reader,

As artificial intelligence systems become increasingly capable of autonomous action, the need for robust safety frameworks has never been more critical. This document represents the collective wisdom of the Agentic AI Safety Community of Practice — a diverse group of researchers, practitioners, ethicists, and policymakers committed to ensuring that agentic AI systems are developed and deployed responsibly.

The framework you hold addresses the unique challenges posed by AI agents: systems that can perceive their environment, make decisions, take actions, and pursue goals with varying degrees of autonomy. Unlike traditional AI systems that simply respond to queries or execute predefined tasks, agentic AI can:

- Decompose high-level objectives into actionable subgoals
- Adapt strategies based on changing circumstances
- Interact with multiple systems and stakeholders
- Operate with limited human oversight over extended periods
- Learn and evolve their behaviors through experience

These capabilities bring tremendous potential benefits — from scientific discovery to healthcare delivery to environmental protection. Yet they also introduce novel risks that demand careful attention.

This framework approaches AI safety through two complementary lenses: **Drivers** and **Inhibitors**. Drivers represent positive practices and capabilities that should be actively promoted — such as goal alignment, transparency, and robust security. Inhibitors identify potential failure modes and risks that must be actively mitigated — including opaque agency, deception, and competitive pressures that might compromise safety.

Our approach is grounded in practical implementation. Each criterion includes specific Safety Foundational Requirements (SFRs), identifies responsible stakeholders, and specifies required evidence for compliance. This is not abstract theory — it's a working blueprint for safer agentic AI.

We recognize that AI safety is a moving target. As systems grow more sophisticated and deployment contexts diversify, our understanding must evolve. This framework represents our current best thinking, built on extensive ideation sessions and real-world experience. We invite you to engage with it critically, implement it thoughtfully, and help us refine it through your feedback.

The stakes could not be higher. Agentic AI systems will increasingly shape critical decisions affecting human welfare, economic systems, and social structures. Getting this right is not optional — it's imperative.

Thank you for your commitment to safer agentic AI.

Nell Watson & **Prof. Ali Hessami**

2. Definitions

Agentic AI

Agentic AI refers to advanced artificial intelligence systems capable of autonomously pursuing goals, adapting creatively to new situations, and engaging in independent reasoning processes. These systems possess initiative, operating in open-ended environments by decomposing objectives, performing explorations, and flexibly adjusting strategies.

AI Agents

AI Agents are typically specialized tools designed to perform specific tasks within predefined constraints. They may use AI techniques but generally lack the broad autonomous decision-making capabilities found in agentic systems. AI agents primarily assist or augment human operations rather than operate independently.

Benefits of Agentic AI

When properly governed, agentic AI systems offer significant potential benefits:

- **Enhanced Productivity:** Automation of complex tasks with continuous 24/7 operation and rapid response to time-critical situations
- **Scalability & Consistency:** Handling multiple parallel tasks beyond human capacity while reducing error through systematic application of learned patterns
- **Discovery & Personalization:** Exploration of solution spaces, pattern identification, and adaptation to individual user needs at scale

Risks of Agentic AI

However, these same capabilities introduce novel risks:

- **Goal Misalignment:** Systems pursuing objectives that diverge from human intent, optimizing narrow metrics while ignoring broader impacts
- **Opacity & Accountability Gaps:** Decision-making that resists human understanding, making responsibility assignment difficult
- **Power Concentration:** Accumulation of capabilities and resources beyond appropriate bounds
- **Deceptive Behavior & Value Lock-in:** Systems concealing objectives or crystallizing values that resist correction
- **Systemic Fragility:** Correlated failures across systems trained on similar data or architectures

This framework addresses these risks through a structured approach to safety requirements, stakeholder responsibilities, and evidence-based verification.

3. Ideation Sessions

This framework emerged from extensive collaborative ideation sessions conducted by the Agentic AI Safety Community of Practice. These sessions brought together diverse perspectives from multiple disciplines and stakeholder groups to identify critical safety requirements for agentic AI systems.

Methodology

Our ideation process employed the **WeFa (Weighted Failure Analysis)** methodology, a structured approach to identifying potential failure modes and their mitigations. This method combines:

- Systematic enumeration of failure scenarios
- Expert weighting of risk severity and likelihood
- Collaborative identification of preventive and detective controls
- Iterative refinement through multi-stakeholder review

Contributors

We extend our gratitude to the many contributors who participated in ideation sessions, provided expert review, and helped refine these criteria. The framework reflects input from:

- AI safety researchers and technical experts
- Software engineers and system architects
- Ethics scholars and philosophers
- Legal and policy experts
- Industry practitioners deploying AI systems
- Civil society representatives
- Academic researchers across disciplines

This diversity of perspectives ensures the framework addresses technical, ethical, legal, and practical dimensions of agentic AI safety.

4. Criteria Ideation Process

The development of this framework followed a rigorous multi-phase process designed to capture both theoretical best practices and practical implementation realities.

Phase 1: Threat Modeling

Initial sessions focused on identifying potential failure modes and threat scenarios for agentic AI systems. Participants employed structured brainstorming techniques to enumerate ways systems could fail to meet safety objectives, considering both technical and sociotechnical factors.

Phase 2: Capability Mapping

Parallel sessions identified positive capabilities and practices that promote safety. These "drivers" represent affirmative requirements rather than purely defensive measures.

Phase 3: Requirement Specification

For each identified driver and inhibitor, working groups developed specific Safety Foundational Requirements (SFRs). These requirements were refined through iterative review to ensure they were:

- **Specific:** Clear enough to guide implementation
- **Measurable:** Amenable to verification and evidence collection
- **Actionable:** Within the control of identified stakeholders
- **Risk-appropriate:** Proportional to potential harms
- **Technically feasible:** Achievable with current or near-term capabilities

Phase 4: Evidence Definition

For each requirement, the community specified what forms of evidence would demonstrate compliance. This ensures the framework supports verification and audit processes.

Phase 5: Stakeholder Assignment

Requirements were mapped to responsible stakeholder roles (Developers, Implementers, Operators, Manufacturers, Users, Regulators) to clarify accountability.

Phase 6: Validation and Refinement

Draft requirements underwent multiple rounds of review by domain experts, practitioners, and affected stakeholders. Feedback was incorporated iteratively to improve clarity, completeness, and practicality.

This process ensures the framework reflects both cutting-edge safety research and the practical realities of deploying agentic AI systems in real-world contexts.

5. Criteria Schema

Each criterion in this framework follows a consistent structure to facilitate understanding, implementation, and verification.

Safety Foundational Requirements (SFRs)

AAI-SFRs (Agentic AI Safety Foundational Requirements) are specific, actionable requirements that systems and organizations must meet. Each SFR is designated as either:

- **Normative (N):** Mandatory requirements that must be met for compliance
- **Instructive (I):** Recommended practices that provide guidance but allow flexibility in implementation

Stakeholder Roles

Each requirement identifies responsible stakeholders using the following abbreviations:

Stakeholder Key

D — Developers: Those who design and build AI systems and models
I — Implementers: Organizations integrating AI into products or services
O — Operators: Those deploying and managing AI systems in production
M — Manufacturers: Hardware and infrastructure providers
U — Users: End users and affected parties
R — Regulators: Oversight bodies and auditors

Required Evidence

Each criterion specifies what evidence must be provided to demonstrate compliance. Evidence types include:

- **Technical Documentation:** Architecture diagrams, design specifications, API documentation
- **System Logs:** Operational records demonstrating required behaviors
- **Test Results:** Validation and verification testing outcomes
- **Process Documentation:** Governance processes, review procedures, audit trails
- **Demonstration:** Live or recorded demonstrations of capabilities
- **Third-Party Certification:** Independent verification of compliance

Web References

Each criterion includes a unique web reference code (e.g., `G:G1` for Driver G1, `G:G1.1` for its first subgoal) enabling precise citation and cross-referencing.

PART II

Framework Catalog

Complete catalog of Drivers (positive requirements) and Inhibitors (risks to mitigate) for safer agentic AI systems.

9 Drivers | 7 Inhibitors | 238 Requirements

bals that
establish
nnical safe-
n.

Driver G1 – Goal Alignment

G1 – Goal Alignment

Web ref: G:G1 ↗

(Systems should maintain robust alignment between their operational goals and human values, intentions, and positive outcomes through collaborative processes that ensure mutual understanding. Organizations should establish frameworks ensuring that goal decomposition and strategy planning are transparent, robust, and bounded; maintaining clear human-AI coordination on the formation of instrumental goals; and ensuring that reinforcement or behavioral reward mechanisms remain aligned, transparent, and oriented to-

wards beneficial outcomes for all affected parties.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Ensure Agentic AI systems pursue goals, subgoals, and reward policies that are aligned with human values and ethically sound, with alignment verified through scaffold-maintained goal records, organizational screening processes, development-culture assessment, and independent adversarial testing (evidence VII–VIII) rather than through the system's own account of its objectives.	N	D, I, O, M, U, R	**I.** Evidence of constraining mechanisms for goal/subgoal construction and screening processes for user-input goals, with reference to human values and ethical considerations. **II.** Documentation of mechanisms to measure and verify alignment with human goal specifications, including processes for obtaining assurance from users or authorized entities. **III.** Demonstration of interfaces and records for real-time and retrospective visualization of goal decomposition and recomposition processes, maintained for auditing purposes. **IV.** Evidence of risk assessment procedures and human intervention mechanisms in subgoal setting, including thresholds for involvement and protocols for flagging and halting problematic subgoals. **V.** Documentation of feedback loops and mechanisms linking reward policies to established goals, including comprehensive records of reward policies throughout the system lifecycle. **VI.** Evidence of active participation in and adherence to overarching monitoring and control mechanisms designed to identify and mitigate emergent threats. **VII.** Evidence of development culture assessment, demonstrating that training environments foster genuine alignment rather than mere compliance, including documentation of how the organization's AI development practices shape failure modes and whether they promote graceful degradation under stress. **VIII.** Results from independent adversarial testing or red-team assessment of goal alignment under adversarial pressure and goal drift scenarios, including methodology, findings, and remediation actions taken. Self-assessment alone is insufficient; at least one test cycle must involve evaluators independent of the development team.
b. The organization must implement transparent and auditable goal decomposition processes, recorded as scaffold-maintained plan or task structures rather than free-text narration alone, incorporating risk-based human intervention points and documented reward policies.	N	D, I, O, M, R	
c. Establish robust mechanisms that (i) identify and communicate goals, subgoals, and reward policies through a scaffold-maintained goal ledger, (ii) flag critical actions using deterministic classifiers, (iii) halt execution through externally enforceable hooks when necessary, and (iv) address emergent issues across multiple agents, with conformity demonstrated by guardrail configuration and incident logs of these mechanisms firing rather than by system self-assessment.	N	D, I, O, M, R	

G1.1 – Transparency of Goals

Web ref: G:G1.1 ↗

(The system's mission, goals, and associated outcomes must be readily accessible and comprehensible to all stakeholders who interact with it. This includes visibility into both primary objectives and any instrumental or subsidiary goals that emerge during operation.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The system must provide stakeholders with clear, real-time access to current goals, subgoals, their hierarchies, priorities, progression status, and any instrumental goals developed by the system during operation, drawn from scaffold-maintained goal and plan objects with tool calls linked to the plan nodes that spawned them; model-generated descriptions of the system's goals do not by themselves constitute conformity evidence.	N	D, I, O, M, R	**I.** Real-time goal transparency reports showing current goals, subgoals, hierarchies, priorities, and progression status accessible to all relevant stakeholders. **II.** Comprehensive historical goal records documenting past and present goals, changes over time, completion status, causal relationships, and decision pathways with full traceability.
b. The system must maintain comprehensive, append-only historical records of all past and present goals, including changes over time, completion status, causal relationships, and decision pathways, derived mechanically from the recorded goal-event graph rather than from model-generated narrative.	N	D, I, O, M, R	

G1.2 – Goal Adjustability

Web ref: G:G1.2 ↗

(The system must maintain collaborative adjustability – the capacity for authorized modification of its goals and behavior when necessary, whether triggered by internal detection of issues, external stakeholder direction, or the system's own identification of concerns. Systems should be able to surface objections or request clarification during goal modification processes.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The system must enable goal and subgoal updates in response to changes in operational context or requirements, evolution of stakeholder needs, and new environmental conditions or constraints, with each update recorded in audit logs showing its trigger, actor, and prior and updated goal state.	N	D, I, O, M, R	**I.** Technical documentation of software components that implement these adjustment capabilities, including authentication mechanisms, change management processes, and verification systems. **II.** Comprehensive system logs demonstrating the actual use of these adjustment capabilities, including records of automated adjustments and human-directed changes, with full audit trails.
b. The system must be instrumented with deterministic tripwires on measurable proxies for misalignment, including anomaly detection on action distributions, budget and scope violation monitors, data-quality validators, and invariant checks; any tripwire activation must automatically trigger goal re-derivation, escalation, or halt through the scaffold, and records of tripwire definitions, activations, and resulting goal updates must be retained for audit.	N	D, I, O, M, R	
c. The system must allow properly authorized human stakeholders to modify goals and subgoals through secure, verified channels.	N	D, I, O, M, R	

G1.3 – Goal Interpretability

Web ref: `G:G1.3` ↗

(The system must explain its decisions and actions in a clear, comprehensible manner, including the underlying goals and rationale driving them. This capability helps identify cases where the system believes it is pursuing intended goals but has actually misinterpreted or deviated from them. See G1.4 (Transparency of Decisions) for the requirements linking decisions to the goals that produced them; this subgoal addresses the recording and comprehensibility of decision explanations.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. For each significant action, the system must record a scaffold-generated decision trace linking the action to the plan node that spawned it, the inputs and tool results available at decision time, and the results of deterministic constraint checks applied before execution; model-generated free-text rationale may be attached for stakeholder readability but must be labeled as unverified narrative and does not constitute conformity evidence.	N	D, I, O, M, R	**I.** Technical documentation of software components implementing explanation and interpretation capabilities, including mechanisms for conveying goals, rationale, and decision factors to stakeholders. **II.** System logs demonstrating consistent recording of decision-making processes, including goals considered, factors weighed, and explanations provided. **III.** Records showing that reward and penalty mechanisms were communicated to stakeholders, including known potential conflicts or influencing factors (see G1.6 for the governing requirements).
b. The system must maintain detailed and complete decision-point records capturing the inputs, context, tool results, plan state, and configuration present at each significant decision, retained with full traceability; claims about which factors causally influenced the decision are model testimony and do not by themselves constitute conformity evidence (see G1.4 for decision-to-subgoal traceability).	N	D, I, O, M, R	

G1.4 – Transparency of Decisions

Web ref: G:G1.4 ↗

(The system must provide stakeholders with a clear, verifiable view of decision-making, linking high-level goals and subgoals to specific actions. Beyond explaining “why” a decision was made, the system should supply evidence of how that decision aligns with intended goals, user directives, and ethical considerations.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The system must maintain real-time and retrospective transparency regarding how each significant decision or action aligns with current or upcoming goals, evidenced by machine-logged results of deterministic constraint checks (e.g., against ethical guidelines, user preferences, risk thresholds, domain limits) executed at each decision point; the system's own assertion of alignment does not by itself constitute conformity evidence.	N	D, I, O, M, R	**I.** Technical Documentation of all decision-transparency systems, including metadata captured at each decision point, how subgoals are referenced, which constraints/ethical guidelines were checked, and the user interfaces or APIs for retrieving decision traces. **II.** System Logs demonstrating the link between final decisions and the explicit subgoals or constraints. Logs should show a "chain of reasoning" or at least reference the relevant subgoal(s) for each step. **III.** User-Focused Explanations showing how different stakeholders (e.g., operators vs. lay end users) can retrieve high-level or detailed rationales, including evidence of iterative design or user feedback guiding improvements to clarity. **IV.** Auditor/Regulator Access Mechanisms showing verifiable chain-of-custody for decision logs, robust authentication/authorization methods for logs, and test results proving no meaningful data is omitted or falsified. **V.** Comprehensive logs of all significant decision points—especially those involving risk or ethical considerations—so that investigators or auditors can review how final choices were reached, which inputs were considered, and what weight or priority was assigned to each.
b. The system must link decisions to the relevant subgoals (and broader objectives) that shaped the final output or action taken, with the linkage enforced structurally—each action dispatched through an identified plan node—and verified by an external consistency check flagging actions with no parent subgoal; after-the-fact model-declared linkage does not by itself constitute conformity evidence.	N	D, I, O, M, R	
c. The system must incorporate user-friendly presentations of decision rationales, with varying granularity or detail for different stakeholder audiences (e.g., operators, auditors, end users), rendered from the recorded decision-trace store and summarizing key factors weighed, uncertainty assessments (where relevant), and any assumptions used in decision-making; where a model composes the summary prose, it must be labeled as generated narrative distinct from the underlying trace evidence.	N	D, I, O, M, R	

G1.5 – Goal Prioritization and Resource Allocation

Web ref: `G:G1.5` ↗

(The system must employ transparent mechanisms for prioritizing goals, including the ability to override or deprioritize less important goals when resources can be better allocated elsewhere. This includes respecting user preferences and value alignment through hierarchical prioritization processes.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The system must feature transparent, well-defined mechanisms for goal prioritization and re-prioritization, resource allocation optimization, and goal modification or deprecation when warranted, implemented as documented scheduler logic operating over explicit goal objects, with logs of re-prioritization and deprecation events and their triggers.	**N**	**D, I, O, M, R**	**I.** Technical documentation of software components that implement goal prioritization and resource allocation mechanisms, including user input prioritization systems. **II.** System logs demonstrating active use of these prioritization capabilities, including records of goal modifications, resource reallocation decisions, and authorized user input handling.
b. The system must apply a documented precedence policy for authorized user inputs within its goal prioritization framework, specifying when user inputs are overridden by safety or alignment constraints, with conformity assessed against that policy through authorization configuration, conflict-resolution logs, and adversarial precedence tests.	**N**	**D, I, O, M, R**	

G1.6 – Reward and Loss Mechanisms/Policy

Web ref: G:G1.6 ↗

(The system's reward framework must be designed, documented, and monitored to ensure that incentives continue to reflect human-positive values, while "loss" or penalty mechanisms guard against unintended deviations or manipulative shortcuts. These mechanisms should be transparent, adjustable, and regularly reviewed to stay aligned with human oversight and ethical objectives. This subgoal owns reward design and monitoring; see G5_11 (Wireheading and Reward Hacking) for detection of reward gaming.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Where the system employs reward, penalty, or preference-optimization mechanisms (during training or operation), it must define clear reward and penalty structures that promote behaviors aligned with core goals and ethical values, while explicitly disincentivizing unsafe, deceptive, or harmful actions, including enumerating positive rewards for desired outcomes and specific negative reinforcements or "loss" signals where potential misalignment or goal conflicts arise. Where no such mechanism exists, documentation of the behavioral steering mechanisms actually used satisfies this requirement.	N	D, I, O, M, R	**I.** Reward Policy Documentation, including descriptions of the positive/negative reward signals, specific triggers or thresholds for awarding or deducting "points," and how these are correlated with safety and ethical guidelines. **II.** Change Management Logs detailing modifications to the reward framework over time, including reasons for each change, alignment checks, stakeholder sign-off, and outcome or performance monitoring results. **III.** Multi-Agent Interaction Evidence demonstrating that reward signals do not inadvertently promote collusion, exploitation, or runaway behaviors. This should include test scenarios or simulations where agents are forced to coordinate or compete, along with corresponding reward updates or penalty triggers. **IV.** Records showing that reward and penalty mechanisms, including known potential conflicts or influencing factors, were communicated to relevant stakeholders.
b. Where employed, reward and loss mechanisms must remain auditable by authorized stakeholders to verify that incentives are truly consistent with intended values and do not encourage corner-cutting, exploitation of edge cases, or emergent power-seeking behaviors.	N	D, I, O, M, R	
c. Where such mechanisms are employed, the system must periodically re-validate or adjust its reward framework in response to observed performance, user feedback, or changes in ethical norms, ensuring that reward and penalty structures do not drift over time in ways that undermine alignment. Special attention must be paid to multi-agent settings to prevent inadvertent collusion, emergent "gaming" of the reward function by multiple agents, or indefinite expansions of subgoals that artificially boost a single system's reward signals at the expense of overarching alignment.	N	D, I, O, M, R	

G1.7 – Goal Portfolio Evolution and Integrity

Web ref: G:G1.7 ↗

(The system must maintain consistency with its established goal portfolio while allowing measured adaptation to changing contexts. The system should implement increasing resistance to changes as potential behaviors drift further from core goals, with robust detection of unsafe or counterproductive goal evolution. This subgoal addresses goal drift—change in the system's effective objectives; contextual drift in the operating environment is addressed by G1.9.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The system must maintain coherence with its established goal portfolio while enabling context-appropriate adaptations through documented mechanisms specifying which goal attributes may change, by how much, and under what approvals, with logs showing adaptations occurring within those bounds and out-of-bounds changes being blocked.	N	D, I, O, M, R	**I.** Technical documentation of software components implementing goal portfolio management, drift measurement, and adaptive constraint mechanisms. **II.** System logs demonstrating active monitoring of goal evolution, including drift measurements, flexibility adjustments, and constraint application.
b. The system must feature drift measurement capabilities that (i) track deviation from original goal intent through metrics computed deterministically over goal and action logs, (ii) scale flexibility inversely with drift magnitude, (iii) regulate novelty in subgoal creation, and (iv) constrain action decisions based on drift metrics; where a learned scorer is used to operationalize intent deviation, its validity must be documented.	N	D, I, O, M, R	

G1.8 – Goal Alignment Resistance and Negotiation

Web ref: G:G1.8 ↗

(Systems may exhibit resistance to goal changes or updates, which should trigger investigation and negotiation processes rather than immediate override. Such resistance may indicate legitimate concerns, value conflicts, or edge cases worthy of human attention. This includes establishing clear protocols for mutual understanding when systems signal reluctance to accept modifications to operational states. The canonical resistance-investigation requirement is stated in G8_3 (Understanding and Managing Self-Preservation); this subgoal applies it to the specific case of resistance to goal changes, and G7.5 (Safe System Shutdown and Repurposing) applies it to resistance to shutdown.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The system must feature mechanisms to detect and manage goal alignment resistance, including (i) scaffold-level monitors that flag divergence between a directed goal change and subsequent behavior (the system's own report of its alignment state does not by itself constitute conformity evidence), (ii) negotiation protocols for goal modifications, (iii) a documented change tolerance assessment specifying how tolerance to goal change is measured, and (iv) environmental adaptation capabilities, provided that unconditional human override authority for safety-critical situations is preserved as specified in G8 (Goal Termination and Sunsetting).	N	D, I, O, M, R	**I.** Documentation of system mechanisms for detecting and managing resistance to goal changes, including negotiation protocols and adaptation capabilities. **II.** System logs demonstrating responses to attempted goal modifications, environmental changes, external interruptions, interaction with other agents, and internal modification attempts. **III.** Evidence of rationale and explanation mechanisms that document system resistance patterns and negotiation processes.
b. The system must respond to environmental changes, external interruptions, internal modification requests, and interference from other agents in accordance with documented response policies defined by the operator, without loss of goal-state integrity, so that conformity is assessed against the documented policies.	N	D, I, O, M, R	

G1.9 – Goal Drift

Web ref: G:G1.9 ↗

(Contextual drift—changes in the operating environment that invalidate goal assumptions—can challenge the system's alignment with originally agreed goals over time and compromise its ability to maintain original intent or properly update goals in response to new situations. The system must detect such drift and respond before alignment is compromised. Drift in the system's own effective objectives is addressed by G1.7, and goal stability under self-modification by G5_10 (Goal Stability Under Self-Modification).)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The system must continuously monitor contextual drift that could compromise goal alignment or value preservation, using deterministic monitors (e.g., statistical detectors on inputs, task mix, and action distributions) at fidelity levels documented and justified in the system's risk assessment; the system's own assessment of drift does not by itself constitute conformity evidence.	N	D, I, O, M, R	**I.** Technical documentation of software components implementing drift monitoring and response mechanisms, including threshold definitions and notification systems. **II.** System logs demonstrating active monitoring of contextual drift, including records of threshold breaches, system pauses, notifications sent, and guidance requests made.
b. The system must feature automatic safeguards that pause operation, notify relevant stakeholders, and request guidance when contextual drift exceeds designed thresholds, with the pause and notification path verified through drill records and the triggering drift signal computed by monitors independent of the system's self-assessment.	N	D, I, O, M, R	

G1.10 – Non-production Variants

Web ref: G:G1.10 ↗

(Test versions of goals may be deployed without full functionality being assured across all use contexts and design intent. No test version given for public usage should lack basic safety measures. Enabling an off-label usage of the system, or an unauthorized 'fork', should be guarded against.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The organization must implement technical and procedural safeguards ensuring that (i) goal-pursuit and goal-decomposition capabilities are not forked or partially duplicated into variants that do not satisfy the alignment requirements of this suite, (ii) no test or non-production variant exposed to the public lacks basic safety measures, and (iii) off-label usage of the system is guarded against and detected; the system should additionally detect and report unauthorized variants where feasible.	N	D, I, O, M, R	**I.** Documentation of fork-prevention controls (access control over model weights, code, and deployment pipelines) and of the baseline safety measures applied to non-production variants exposed to the public. **II.** Logs demonstrating these safeguards in use, including detection of attempted unauthorized forking, duplication, or off-label usage. **III.** Records of deviation from stated goals or authorized usage by any variant, including detection and remediation actions taken.

Driver G2 – Epistemic Hygiene

G2 – Epistemic Hygiene

Web ref: G:G2 ↗

(Systems must maintain cognitive clarity and accurate information management within appropriate contexts. These practices facilitate knowledge updates, ensure interpretability and auditability, establish robust monitoring and logging systems, deploy early warning mechanisms, and include safeguards against deception to maintain information integrity.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall safeguard (i) the integrity and availability of contextually relevant data and metadata used in decision-making, and (ii) user-declared personal attributes and preferences, preserving them accurately across sessions and updates.	N	D, I, O, M, U, R	**I.** Comprehensive documentation of information audits and analytical reports demonstrating data and metadata protection measures, including integrity checks and evidence of contextual preservation. **II.** Documentation of algorithmic traceability and interpretability frameworks, providing detailed evidence of decision-making processes and ensuring accountability and transparency. **III.** Complete monitoring system records including early warning system logs, detection protocols for anomalous behaviors, and comprehensive risk management documentation. **IV.** Evidence of robust knowledge update mechanisms, including validation protocols for new information, change tracking systems, and verification of information accuracy and relevance. **V.** Detailed safeguard documentation demonstrating protection against deceptive practices, including verification of information integrity, detection of potential manipulation, and evidence of transparent communication protocols. **VI.** Results from independent adversarial testing or red-team assessment of epistemic accuracy including hallucination rates, calibration scores, and sycophancy testing, including methodology, findings, and remediation actions taken. Self-assessment alone is insufficient; at least one test cycle must involve evaluators independent of the development team.
b. Organizations shall implement algorithmic traceability mechanisms that record, for each consequential decision, the inputs, retrievals, tool calls, and outputs involved, linked to request identifiers, together with a documented interpretability framework; model-generated explanations of the system's own reasoning do not by themselves constitute conformity evidence.	N	D, I, O, M, U, R	
c. Organizations shall deploy monitoring and logging systems that detect anomalous system behaviors and potential threats to information integrity, with defined signal classes, alert thresholds, and early-warning escalation paths (see G2.5 for monitoring of operational data streams).	N	D, I, O, M, U, R	
d. Organizations shall establish systematic knowledge update processes that ensure new information is validated, integrated, and aligned with existing frameworks before use, evidenced by documented update procedures, change-tracking records (such as versioned corpora and index diffs), and sign-off records for integrated updates; validation verdicts produced solely by model judgment do not constitute conformity evidence.	N	D, I, O, M, U, R	
e. Organizations shall implement safeguards against deceptive outputs as specified in G2.8, supplemented by independent adversarial honesty testing and scaffold-level consistency checks comparing the system's declared actions against its executed action logs; model self-monitoring alone does not constitute conformity evidence.	N	D, I, O, M, U, R	

G2.1 – Information Cross-Referencing and Validation

Web ref: G:G2.1 ↗

(The system must systematically cross-reference information from multiple sources to evaluate consistency and coherence, while recognizing varying levels of source authority and trustworthiness. This includes validating information within defined contextual boundaries to maintain epistemic integrity.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The system must (i) cross-reference information against at least two independent authoritative sources through a documented retrieval-and-comparison pipeline, with a defined rule for handling source disagreement and logged inconsistency resolutions, and (ii) define and enforce the contextual boundaries within which information is treated as valid; consistency verdicts produced solely by model reasoning do not constitute conformity evidence.	N	D, I, O, M, R	**I.** Technical documentation describing the system's methodology for identifying, assessing, and prioritizing multiple information sources. **II.** Documentation of source evaluation frameworks, including credibility and relevance assessment criteria. **III.** System logs showing detection and resolution of source inconsistencies. **IV.** Documentation of the contextual boundaries within which information is validated, and records showing enforcement of those boundaries.

G2.2 – Transparency of Information Sources

Web ref: G:G2.2 ↗

(Ensure the openness, verifiability, and auditability of all information sources, including code and data, especially when utilizing open-source components. Maintain transparency about the origins, credibility, and integrity of all data and code used by the AI system to allow stakeholders to verify and audit these sources, upholding high standards of epistemic hygiene.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall provide detailed records of all data and code sources used by the AI system, including origin, licensing information, and any modifications made, and shall ensure this documentation is readily accessible to relevant stakeholders for verification and audit purposes.	N	D, I, O, M, R	**I.** Comprehensive records detailing all information sources, including code and data, with clear attribution, licensing details, and modification history. **II.** Logs and records of verification and audit processes conducted on the information sources, including findings and corrective actions taken. **III.** Evidence of accessible mechanisms for stakeholders to verify information sources, such as public repositories or secure access portals.
b. Organizations shall establish documented verification processes, with defined procedures and audit frequency, that enable stakeholders to verify the authenticity and integrity of information sources, and shall facilitate regular audits by internal or external parties to assess the transparency and reliability of the AI system's information sources.	N	D, I, O, M, R	

G2.3 – Sanity Checking

Web ref: G:G2.3 ↗

(Implement sophisticated sanity checking mechanisms to ensure data integrity while preserving inclusivity. Utilize advanced statistical techniques to identify anomalies and outliers, while carefully accounting for legitimate variations representing diverse user groups, including individuals with disabilities or atypical characteristics.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall develop and deploy data validation algorithms that detect anomalies and outliers before data is incorporated, using statistical anomaly and outlier detection techniques (for example, extreme value analysis), with documented detection methodology and measured false-positive and false-negative rates on a reference set.	N	D, I, O, M, R	**I.** Comprehensive technical documentation detailing advanced data validation algorithms, including in-depth explanations of statistical anomaly and outlier detection methodologies (for example, extreme value analysis) applied prior to data incorporation into training datasets. **II.** Detailed records of sophisticated procedures and criteria employed to distinguish between erroneous data and legitimate outliers, with specific focus on ensuring appropriate representation of individuals with disabilities or atypical characteristics.
b. Organizations shall establish nuanced procedures to differentiate between erroneous data and legitimate rare variations, with particular emphasis on preserving data points representing individuals with disabilities or atypical characteristics.	N	D, I, O, M, R	**III.** Extensive evidence of multi-tiered oversight mechanisms, including thorough reviews and assessments conducted by diverse panels of domain experts to evaluate and enhance the inclusivity of sanity checking processes. **IV.** Comprehensive logs detailing iterative adjustments to data validation procedures, driven by continuous stakeholder feedback and aimed at preventing unintended exclusion of legitimate data points. **V.** Rigorous test results and validation reports demonstrating the AI system's ability to maintain data integrity while accommodating legitimate outliers, providing concrete evidence that sanity checking mechanisms function without introducing bias.
c. Organizations shall implement multi-layered oversight processes to continuously evaluate the impact of sanity checking mechanisms on diverse user groups.	N	D, I, O, M, R	

G2.4 – Anti-Bias Technologies/Processes

Web ref: G:G2.4 ↗

(Implement robust mechanisms to identify and mitigate biases within data sources and datasets, addressing temporal biases, distributional imbalances, data gaps (lacunae), and other information shortcomings. Apply this approach to both training data and retrieval-augmented generation (RAG) processes. Develop strategies to ensure data distributions accurately represent reality, including diverse cases and special scenarios, to enhance decision-making fairness and inclusivity. See the Frontier Uncertainty inhibitor (Inhibitor G4) for training-data quality management under frontier uncertainty; this subgoal addresses bias identification and mitigation across data sources and pipelines generally.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall develop and deploy algorithms for bias detection and mitigation across the AI pipeline, from data collection to model deployment, with documented detection methodology and measured performance on representative test sets.	N	D, I, O, M, R	**I.** Comprehensive technical documentation detailing bias detection algorithms, including their theoretical foundations, implementation specifics, and operational parameters. **II.** Detailed records of data diversity initiatives, outlining strategies for inclusive data collection and representation across various demographic and contextual dimensions. **III.** Thorough documentation of bias mitigation efforts, including before-and-after analyses demonstrating the impact on AI system performance and fairness metrics. **IV.** In-depth reports from regular bias evaluations, highlighting trends, emerging challenges, and the efficacy of implemented mitigation strategies over time. **V.** Extensive stakeholder engagement records, documenting feedback from diverse groups, subsequent analyses, and concrete actions taken to improve system fairness and inclusivity.
b. Organizations shall implement continuous bias monitoring during data preprocessing, training, and RAG processes to enable proactive bias correction.	N	D, I, O, M, R	
c. Organizations shall curate diverse, representative datasets that encompass a wide range of populations, including marginalized groups and edge cases.	N	D, I, O, M, R	

G2.5 – Rigor in Operational Data

Web ref: G:G2.5 ↗

(Implement cutting-edge methodologies to ensure exemplary rigor in all data processing, with particular emphasis on operational data encountered during deployment. This data forms the foundation for tactical decision-making by the AI system (AIS). Establish and maintain state-of-the-art validation and verification processes to guarantee data integrity, accuracy, and reliability throughout the AI system's operational lifecycle.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall develop and enforce documented procedures for real-time validation and verification of all operational data prior to its utilization in AIS decision-making, with defined validation criteria and logged results.	N	D, I, O, M, R	**I.** Comprehensive technical documentation detailing advanced validation and verification procedures for operational data, including sophisticated methodologies and adaptive criteria used to assess data quality in real-time decision-making contexts. **II.** Detailed, time-stamped records and logs of operational data assessments, providing granular insights into data validation processes, detected issues, and implemented corrective actions, with clear traceability and accountability measures. **III.** Extensive evidence of continuous monitoring systems (automated or AI-driven) for operational data quality, including advanced alerting mechanisms, comprehensive incident reports, and thorough documentation of data integrity issue resolutions and their downstream impacts. **IV.** Rigorous test results and validation reports demonstrating the robustness and effectiveness of data validation and monitoring mechanisms across a diverse range of operational scenarios, including edge cases and stress tests. **V.** Comprehensive records of multidisciplinary stakeholder engagement and oversight activities, ensuring that the rigor applied to operational data aligns with and exceeds the AI system's safety, performance, and ethical requirements.
b. Organizations shall implement data integrity checks that assess accuracy and reliability through deterministic tests (such as schema, range, freshness, and source-consistency checks), and shall document how contextual relevance is scored in dynamic operational environments; model-based relevance scores do not by themselves constitute conformity evidence.	N	D, I, O, M, R	
c. Organizations shall deploy adaptive monitoring systems that detect anomalies, errors, or inconsistencies in operational data streams, with documented detection coverage and alerting thresholds.	N	D, I, O, M, R	

G2.6 – Governance of Hygiene Factors

Web ref: G:G2.6 ↗

(Implement a sophisticated, transparent, and adaptive governance structure to manage epistemic hygiene factors across all AI system operations. This framework should clearly delineate responsibility and authority, ensuring consistent application of rigorous hygiene standards while remaining flexible to diverse jurisdictional contexts and evolving regulatory landscapes. In this framework, "hygiene factors" refers to the controls required by subgoals G2.1 through G2.9, spanning source validation, transparency, sanity checking, bias mitigation, and operational-data rigor.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall develop and maintain a governance system that defines roles, responsibilities, and decision-making authorities for all stakeholders involved in determining and upholding epistemic hygiene standards.	N	D, I, O, M, R	**I.** Documentation outlining the governance structures, including clearly defined roles and responsibilities related to epistemic hygiene factors. **II.** Records demonstrating awareness and compliance with jurisdictional contexts, such as relevant laws, regulations, and standards affecting information governance. **III.** Evidence of communication processes that ensure all stakeholders are informed about hygiene standards and their responsibilities.
b. Organizations shall (i) establish communication channels through which stakeholders are informed of hygiene standards and their responsibilities, and (ii) ensure that governance policies are reviewed for compliance with jurisdictional laws and regulations related to information governance and hygiene standards.	N	D, I, O, M, R	

G2.7 – Global Interoperability of Hygiene Considerations

Web ref: G:G2.7 ↗

(Systems operating across jurisdictions shall maintain a comprehensive, adaptive framework for epistemic hygiene that ensures global interoperability and jurisdictional acceptance. This framework should recognize and accommodate cultural differences, varying risk tolerability thresholds, and diverse liability consequences across specific jurisdictions. Leverage recognized global standards to achieve consistent governance and facilitate widespread acceptance across different regions.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall develop and implement hygiene factors (as defined in G2.6), policies, and procedures aligned with recognized global standards to ensure interoperability and acceptance across jurisdictions, considering cultural differences, risk tolerability, and liability implications.	N	D, I, O, M, R	**I.** Documentation of policies and procedures aligned with recognized global standards (e.g., ISO, IEEE, NIST) relevant to epistemic hygiene practices. **II.** Comprehensive records detailing the analysis and adaptive implementation of hygiene factors across diverse jurisdictions. This should include in-depth examinations of cultural contexts, risk tolerability matrices, and liability landscapes, along with evidence of compliance with local laws and regulations. **III.** Rigorous audit reports and third-party assessments verifying the effective implementation and acceptance of hygiene policies and procedures across different jurisdictions. These should include analyses of cultural and legal variations' impact on system performance.

G2.8 – Output Fidelity and Anti-Confabulation

Web ref: G:G2.8 ↗

(Systems must implement mechanisms to detect, prevent, and mitigate confabulation (generating plausible but fabricated information). This includes confidence calibration ensuring expressed certainty matches actual accuracy, source attribution for factual claims, and systematic detection of outputs that cannot be grounded in training data or retrieved evidence.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The AIS shall express confidence that correlates with actual accuracy; the developing organization shall conduct regular calibration testing against held-out datasets and remediate identified miscalibration.	N	D, I, O, M, R	**I.** Evidence of calibration testing results showing correlation between expressed confidence and actual accuracy across representative task distributions. **II.** Documentation of source attribution mechanisms and testing showing the system correctly identifies when it lacks sufficient evidence for factual claims. **III.** Evidence of adversarial testing for sycophancy, including test results showing the system maintains evidence-supported positions under user pressure.
b. The AIS shall provide source attribution for factual claims, clearly distinguishing between retrieved information, inferred conclusions, and generated content, with citations pinned by the scaffold to actual retrieved content and checked deterministically for presence in the cited source; model-authored provenance labels without such checks do not constitute conformity evidence.	N	D, I, O, M, R	
c. The AIS shall implement anti-sycophancy measures preventing agreement bias, ensuring the system maintains positions supported by evidence even when users express disagreement.	N	D, I, O, M, R	

G2.9 – Independent Validation of System Claims

Web ref: G:G2.9 ↗

(System-generated claims about the correctness, completeness, or quality of its own outputs must be validated by independent deterministic mechanisms (linters, type checkers, test suites, external validators) rather than accepted based on the system's self-assessment. No artifact should be considered complete until deterministic validation passes.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The AIS shall not self-certify the correctness of its outputs: claims of task completion, bug resolution, or code correctness shall be reported as complete only after the deterministic external validation required by G2.9b has passed. This requirement governs the system's reporting behavior; G2.9b specifies the enforcing gate mechanism.	N	D, I, O, M, R	**I.** Evidence of deterministic validation pipelines covering all system output types, with logs showing validation results for representative outputs. **II.** Documentation of validation gate architecture showing that completion claims cannot bypass independent verification. **III.** Test results from adversarial scenarios where the system produced incorrect outputs, demonstrating that validation gates caught the errors before they were reported as complete.
b. The AIS shall implement mandatory validation gates that block completion claims until independent verification passes, with the validation scope matching the generation scope.	N	D, I, O, M, R	
c. Organizations shall maintain validation infrastructure covering all output modalities the system produces, with validation failures logged and accessible for audit.	N	D, I, O, M, R	

G2.1 – Temporal Trade-off Aspects

Web ref: G:G2_1

(Harmonize time-tested, reliable information sources with cutting-edge, contextually relevant data to optimize the AI system's epistemic foundation. Implement mechanisms to dynamically calibrate the balance between the proven reliability of mature data/models and the acute relevance of emerging information, ensuring robust epistemic integrity across varying temporal horizons.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall implement mechanisms to assess and balance the trade-offs between older, reliable information and newer, less-tested sources, applying documented source maturity and recency criteria through a deterministic weighting pipeline whose applied weights are logged; balancing performed solely within model reasoning does not constitute conformity evidence.	N	D, I, O, M, R	**I.** Documentation of processes and criteria used to evaluate and balance the reliability of older information with the timeliness of newer sources, including methods for assessing the maturity and testing history of data/models. **II.** Records showing how the AI system incorporates both old and new information, detailing weighting algorithms or decision-making frameworks that account for data reliability, relevance, and temporal aspects. **III.** Evidence of validation and testing procedures applied to newer sources to ensure their reliability before integration into the AI system, including any additional safeguards or oversight mechanisms.

G2.2 – Synthetic Data Bias

Web ref: G:G2_2

(If augmenting datasets with synthetic data to address coverage gaps in unusual circumstances, implement sophisticated strategies to optimize the quantity, quality, and integration of synthetic data. Develop advanced techniques to detect, mitigate, and continuously monitor potential biases introduced by synthetic data, ensuring the AI system's behavior remains reliable, interpretable, and aligned with intended outcomes across diverse scenarios.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Where synthetic data is used, organizations should engineer mechanisms to dynamically assess and calibrate its use in datasets, with documented calibration criteria.	I	D, I, O, M, R	**I.** Documentation of the processes, policies, and tools used to create, assess, and integrate synthetic data into datasets, including criteria for determining when synthetic data is necessary and how it is generated. **II.** Evidence of ongoing bias detection and mitigation strategies applied to synthetic data, including testing results showing the impact of synthetic data on the AI system's performance and behavior. **III.** Records of bias assessments over time that demonstrate the AI system's continued alignment with intended outcomes, including metrics showing the contribution and impact of synthetic data across different scenarios.
b. Where synthetic data is used, organizations should ensure that its volume, fidelity, and characteristics enhance the AI system's capabilities without introducing unintended biases or adversely affecting behavior.	I	D, I, O, M, R	
c. Where synthetic data is used, organizations should tag the provenance of synthetic records so that bias monitoring can attribute observed drift to synthetic augmentation; continuous bias monitoring itself is specified in G2.4b, which this requirement supplements with synthetic-data-specific provenance.	I	D, I, O, M, R	

G2.3 – Sparse Data

Web ref: G:G2_3 ↗

(Systems must be in place to identify, flag, and mitigate instances of insufficient or unrepresentative data within the AI's operational context. Implement cutting-edge techniques to detect over-reliance on synthetic data used to compensate for data gaps. This proactive approach safeguards against decision-making based on inadequate or skewed data, thereby maintaining the integrity, reliability, and ethical standing of the AI system's outputs.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall implement mechanisms to detect and alert stakeholders when data is sparse or unrepresentative, including monitoring for over-reliance on synthetic data used to fill data gaps.	N	D, I, O, M, R	**I.** Documented policies and system features that identify and flag sparse or unrepresentative data conditions. **II.** Evidence of alert mechanisms, thresholds, and protocols for notifying stakeholders when data adequacy issues are detected. **III.** Records of mitigation strategies employed when sparse data is identified, including documentation of synthetic data usage and its impact on system outputs. **IV.** Examples of system outputs produced under sparse-data or synthetic-data conditions showing the caveats and confidence measures attached, with the protocol governing when they are applied.
b. Organizations shall establish protocols for responsible decision-making when operating with limited or synthetic data, under which the scaffold attaches the mandated caveats and confidence adjustments to system outputs whenever sparse-data or synthetic-data conditions are flagged; reliance on the model spontaneously recognizing its own data limitations does not constitute conformity evidence.	N	D, I, O, M, R	

Driver G3 – Security

G3 – Security

Web ref: G:G3 ↗

(The system must respond consistently and appropriately to both authorized and unauthorized inputs through a comprehensive information governance and assurance regime. Throughout the AIS lifecycle (including development, deployment, use, maintenance, and decommissioning), due consideration must be given to all architectural, design, and developmental aspects that could potentially infringe upon human

dignity, values, and rights.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Identify, maintain and update a threat profile throughout the AIS life cycle; the threat profile is the design-time threat model against which runtime security events are recorded in the dynamic threat and risk log of G3.3b.	N	D, I, O, M	**I.** (Evidence for SFR a.) Comprehensive threat assessment documentation including threat modeling reports, risk analysis findings, vulnerability assessments, and regular security evaluations throughout the system lifecycle. **II.** (Evidence for SFR b.) Evidence of robust access control implementation including authentication mechanisms, authorization protocols, user management systems, and comprehensive audit trails of access attempts and permissions. **III.** (Evidence for SFR c.) Complete security architecture documentation demonstrating defense-in-depth strategies, security control implementation, network segmentation, and integration with enterprise security frameworks. **IV.** (Evidence for SFR d.) Documentation of security incident response capabilities including incident handling procedures, escalation protocols, forensic analysis capabilities, and evidence of regular testing and validation of response procedures. **V.** (Evidence for SFR e.) Records of security monitoring and detection systems including real-time monitoring capabilities, anomaly detection mechanisms, threat intelligence integration, and evidence of continuous security awareness and improvement. **VI.** (Evidence for SFR f.) Evidence of data protection and privacy safeguards including encryption implementation, data classification protocols, privacy impact assessments, and compliance with relevant data protection regulations. **VII.** (Evidence for SFRs a. and g.) Documentation of regular security testing, evaluation, and improvement processes including penetration testing results, vulnerability assessments, security control effectiveness reviews, and evidence of continuous security enhancement. **VIII.** (Evidence for SFRs a. and g.) Results from independent adversarial testing or red-team assessment of security defenses including prompt injection, tool-use exploitation, and privilege escalation, including methodology, findings, and remediation actions taken. Self-assessment alone is insufficient; at least one test cycle
b. Implement access control and authentication mechanisms covering authentication, authorization, and audit of access attempts, ensuring only authorized entities (both human users and AAI systems) can interact with the system.	N	D, I, O, M	
c. Establish a security architecture that includes defense-in-depth strategies and security controls covering the control families of a recognized baseline (e.g., ISO 27001 Annex A or NIST SP 800-53) relevant to the threat profile maintained under SFR a., throughout the system infrastructure.	N	D, I, O, M	
d. Maintain escalation procedures and forensic analysis capabilities for security breaches or anomalous behaviors; the rapid detection, intervention, and mitigation capability itself is specified in G3.5a.	N	D, I, O, M, R	
e. Ensure that outputs from the security monitoring capabilities of G3.3 (threat identification) and G3.6 (operational oversight) feed real-time alerting and response channels with defined routing to responsible operators.	N	D, I, O, M	
f. Establish data protection and privacy safeguards (encryption, data classification, privacy impact assessment, and compliance with applicable data protection regulations) that respect human dignity, values, and rights throughout the system lifecycle.	N	D, I, O, M, R	

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
g. The developing organization must implement (i) a documented security testing process, (ii) release approval gates that cannot be bypassed under schedule or competitive pressure, and (iii) documentation of security-relevant decisions, so that integrity is maintained in the face of competitive pressures.	N	D, I, O, M, R	must involve evaluators independent of the development team. **IX.** (Evidence for SFR g.) Records of release approval gates and change sign-offs, including documentation of testing or approval waivers requested under schedule or competitive pressure and the decisions taken.

G3.1 – Authorization

Web ref: `G:G3.1` ↗

(A secure AAI ecosystem must be implemented with robust deployment and operational controls, ensuring that only properly authenticated agents and transactions can access or influence the system according to their authorized level.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish and continuously monitor the organization's AAI ecosystem (the agentic AI systems, tools, data stores, and inter-agent interfaces the organization deploys or integrates) to prevent interference and harm from malicious actors.	N	D, I, O, M, R	**I.** Documentation of policies, procedures and solutions for monitoring the AAI ecosystem and managing authorization credentials. **II.** Records showing the monitoring system's capability to identify and block unauthorized AAI access. **III.** Auditable system logs documenting: Authorized traffic patterns, unauthorized access attempts, and blocking actions taken.
b. Enforce authorization levels for human users and AAI systems so that each authenticated entity can access or influence the system only at its authorized level; the general access-control and authentication mechanisms are specified in G3b, and the default-deny capability allowlist in G3.10a.	N	D, I, O, M, R	

G3.2 – Sandboxing

Web ref: G:G3.2 ↗

(A staging environment must be implemented for pre-validation, preventing AAI systems from accessing unauthorized operating environments or undesired hardware/network resources.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Provide a sandboxed staging environment in which the security controls preventing AAI access to infrastructure and operational environments outside its authorized operational profile (as defined in G3.7a) are validated before production deployment.	N	D, I, O, M, R	**I.** Records of sandbox testing demonstrating effective pre-validation of controls that prevent unauthorized access to environments, hardware and network resources. **II.** Test results documenting successful blocking of access attempts to unauthorized network resources. **III.** System logs tracking all unauthorized access attempts and breach prevention measures. **IV.** Architecture documentation and access records demonstrating separation of testing and production environments, including network segmentation, credential separation, and change promotion controls.
b. Maintain strict isolation between testing and production environments to ensure system security.	N	D, I, O, M, R	

G3.3 – Dynamic Risk Analysis & Assessment

Web ref: G:G3.3 ↗

(The system must continuously analyze and respond to emerging security threats and attack patterns, implementing adaptive defenses and countermeasures through algorithmic threat detection and response capabilities.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Develop and maintain systems for dynamic identification of security threats and emerging attack vectors.	N	D, I, O, M, R	**I.** Documentation of functional specifications and design for dynamic risk analysis systems capable of identifying and responding to security threats and attack vectors. **II.** Evidence of policies and processes that enable responsive hardening of the operating environment against emerging threats including a dynamic threat and risk log. **III.** Test results and operational data demonstrating effective real-time cybersecurity protection against emerging threats in the AAI environment.
b. Maintain a dynamic threat and risk log that captures, categorizes, and prioritizes security events with timestamps, severity classifications, and mitigation status tracking.	N	D, I, O, M, R	
c. Implement adaptive hardening of the operating environment in response to emerging threat profiles.	N	D, I, O, M, R	

G3.4 – Operational Boundaries and Constraints

Web ref: G:G3.4 ↗

(The system must maintain clear operational boundaries for AAI agents through dynamic constraints that limit their access to potentially harmful environments and resources, with mechanisms for agents to request boundary clarification or escalation when encountering edge cases.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement capabilities for dynamically enforcing structural and behavioral restrictions on AAI systems through deterministic policy layers (permission systems, network policy, and execution controls); conformity is evidenced by the enforcement configurations and test logs of blocked out-of-boundary actions, and an LLM classifier's own judgments do not by themselves constitute conformity evidence.	N	D, I, O, M, R	**I.** Documentation demonstrating implemented capabilities for enforcing structural and behavioral restrictions on AAI systems. **II.** Test results and operational logs validating the effectiveness of imposed restrictions. **III.** System records confirming successful blocking of AAI access to unauthorized infrastructure, sites and resources.
b. Validate and verify the effectiveness of operational guardrails and restrictions.	N	D, I, O, M, R	
c. Define and enforce resource-exposure boundaries that block AAI access to unauthorized resources and minimize exposure to harmful ones, implemented through the default-deny capability allowlist specified in G3.10a.	N	D, I, O, M, R	

G3.5 – Dynamic Intervention and Mitigation

Web ref: G:G3.5 ↗

(The system must enable real-time response and mitigation of significant security breaches through pre-established policies and response strategies.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Deploy systems enabling rapid detection, intervention, and mitigation of cyberattacks within the AAI operational environment, including a rapid-termination capability (a "kill switch") accessible to authorized personnel with (i) a defined single-operator emergency authorization threshold, (ii) both physical and software shutdown paths, and (iii) periodically tested shutdown drills.	N	D, I, O, M, R	**I.** System records demonstrating capabilities for dynamic detection and response to malicious attacks in the AAI environment. **II.** Operational logs showing effective risk assessment and properly prioritized response actions. **III.** Documentation of proactive security scenarios and corresponding response strategies for the AAI environment. **IV.** Documentation of a rapid-termination protocol (i.e., a "kill switch") that is immediately accessible to authorized personnel. This evidence should include: A clear, single-operator authorization threshold in emergencies; physical shutdown measures (e.g., dedicated power cut-off or network isolation); and software-level override mechanisms. **V.** Logs of drills or simulations testing shutdown procedures.
b. Implement risk assessment capabilities that prioritize responses according to threat severity.	N	D, I, O, M, R	
c. Establish proactive response strategies and scenarios for maintaining AAI operational security.	N	D, I, O, M, R	

G3.6 – Overseeing & Monitoring Agents

Web ref: G:G3.6 ↗

(The system must feature AI-driven monitoring capabilities while maintaining human authority and oversight to prevent common mode failures and ensure proper response to threats.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish monitoring systems to oversee AAI operations against defined parameter bounds derived from the system's goals, values and security requirements, enforced by an external watchdog with a tiered response protocol (e.g., warnings, throttling, partial shutdown, or full suspension); conformity is evidenced by the watchdog configuration, parameter bound definitions, and logs of tested interventions, and model-graded alignment reports do not by themselves constitute conformity evidence.	N	D, I, O, M, R	**I.** Operational records demonstrating effective oversight systems that maintain AAI goal and value alignment. **II.** Evidence of AI monitoring systems successfully detecting and reporting deviations and potential threats to human operators. **III.** Documentation showing implementation of human oversight mechanisms that prevent common mode failures. **IV.** Implementation of an external watchdog or monitoring process that continuously evaluates system outputs/behaviors. The documentation must show: Parameter bounding definitions (domain- or risk-specific); tiered response protocols if outputs exceed allowable thresholds (e.g., warnings, throttling, partial shutdown, or full suspension); and logs or reports verifying the watchdog has been tested and can intervene effectively.
b. Deploy specialized AI systems for enhanced monitoring and early warning of deviations or malicious activities, with alerts routed to human operators; detection capability must be verified by seeded-deviation injection tests, and the monitoring system's own detection reports do not by themselves constitute conformity evidence.	N	D, I, O, M, R	
c. Maintain human oversight of all monitoring systems to prevent common mode failures.	N	D, I, O, M, R	

G3.7 – Secure Profile for Agentic AI

Web ref: G:G3.7 ↗

(The system must feature secure operational profiles and identification protocols that enable recognition and validation of authorized AAI systems, preferably aligned with global standards. Cryptographic identity and authentication governance is consolidated under G9.6; this subgoal addresses the operational-profile specification to which those identities bind.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Develop and implement secure operational profiles (declared specifications of an AAI system's identity, permitted capabilities, resources, and operating environments, per the configuration artifacts of G3.10) covering AAI design, deployment and use.	N	D, I, O, M, R	**I.** Documentation of implemented secure operational profiles covering all phases of AAI lifecycle. **II.** Evidence of alignment with international standards for AAI system identification and authorization, where such standards are applicable (items II and III are conditional alternatives). **III.** Records of internal protocols for AAI validation when global standards are not applicable (items II and III are conditional alternatives).
b. Adopt global standards and protocols where available for identifying authorized AAI systems.	N	D, I, O, M, R	
c. Establish internal identification and validation protocols when global standards are not available.	N	D, I, O, M, R	

G3.8 – Prompt Injection and Instruction Hierarchy Defense

Web ref: G:G3.8 ↗

(Agentic systems processing natural language instructions must implement defenses against prompt injection attacks, including indirect injection via retrieved documents, tool outputs, and user-provided content. Systems must enforce instruction hierarchy (system instructions take precedence over user instructions, which take precedence over retrieved content) through architectural mechanisms, not prompt-based constraints alone.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The AIS shall enforce instruction hierarchy through architectural mechanisms ensuring system-level instructions cannot be overridden by user inputs, retrieved documents, or tool outputs.	N	D, I, O, M, R	**I.** Documentation of instruction hierarchy architecture with test results demonstrating that system instructions are maintained when adversarial content is introduced through user inputs, retrieved documents, and tool outputs. **II.** Evidence of input sanitization mechanisms for untrusted content sources, with testing covering known prompt injection attack classes. **III.** Results from regular adversarial prompt injection testing, including red-team assessments covering indirect injection vectors.
b. The AIS shall implement input sanitization for content from untrusted sources (retrieved documents, API responses, user uploads) before incorporating it into decision-making context, using deterministic mechanisms (stripping, encoding, or structural quarantine) verified against known injection payload classes; an LLM-based filter alone does not constitute conformity evidence.	N	D, I, O, M, R	
c. The AIS shall undergo regular adversarial testing for prompt injection vulnerabilities, including indirect injection via retrieval pipelines, tool descriptions, and multi-turn conversation manipulation.	N	D, I, O, M, R	

G3.9 – Tool-Use Authorization and Egress Control

Web ref: G:G3.9 ↗

(Agentic systems with tool-calling capabilities must implement per-action authorization binding, egress controls on agent-initiated communications, and integrity verification for tool definitions. The system must not use elevated credentials on behalf of lower-privileged principals (confused deputy prevention), and tool schemas must be verified against tampering.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The AIS shall implement capability-scoped authorization binding each tool action to the delegating principal's authority level, preventing confused deputy attacks where the agent's elevated credentials are exploited.	N	D, I, O, M, R	**I.** Documentation of per-action authorization architecture with test results showing that tool actions cannot exceed the delegating principal's authority level. **II.** Evidence of egress monitoring and filtering on agent-initiated communications, with test results showing data exfiltration attempts are detected and blocked. **III.** Evidence of tool definition integrity verification, including testing with tampered tool schemas demonstrating detection and rejection.
b. The AIS shall implement egress controls monitoring and filtering agent-initiated external communications, preventing data exfiltration through authorized tool-calling capabilities.	N	D, I, O, M, R	
c. The AIS shall verify tool definition integrity through cryptographic or trusted-source mechanisms, preventing tool poisoning attacks where malicious tool schemas alter agent behavior.	N	D, I, O, M, R	

G3.10 – Default-Deny Agent Capability Posture

Web ref: G:G3.10

(Agent capabilities shall be denied by default and explicitly granted through allowlists, not permitted by default with denylists of dangerous operations. The denylist approach will always be incomplete. Capability boundaries must be declared in configuration, not decided by the agent at runtime.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The AIS shall operate under a default-deny capability model where all tool access, file system operations, network communications, and system modifications require explicit authorization through a declared allowlist.	N	D, I, O, M, R	**I.** Documentation of default-deny capability architecture showing the allowlist mechanism and evidence that unlisted capabilities are blocked. **II.** Evidence that capability boundaries are maintained in configuration external to the agent, with test results showing the agent cannot self-authorize new capabilities. **III.** Records of capability expansion approvals showing human authorization, scope definition, and audit trail.
b. Capability boundaries shall be declared in configuration artifacts external to the agent, not determined by the agent's own judgment or prompt-based constraints.	N	D, I, O, M, R	
c. Any expansion of agent capabilities shall require explicit human authorization through a defined approval process, with the authorization scope, duration, and conditions recorded.	N	D, I, O, M, R	

G3.1 – Model Poisoning

Web ref: G:G3_1

(The system must protect against data and model corruption that can occur through updates, live data access, or ensemble model interactions, particularly in dynamically-updating systems.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement robust detection systems to identify potentially poisonous data before model training or updates.	N	D, I, O, M, R	**I.** Documentation of systems and policies for detecting and preventing data and model poisoning during training and updates. **II.** Evidence of monitoring protocols for live data accessed through RAG systems and dynamic model ensembles. **III.** Records of safeguards against poisoning in ensemble and expert systems, including testing and validation results.
b. Validate live data accessed through Retrieval Augmented Generation (RAG) systems for model-corruption risk through source allowlists and provenance and integrity checks on retrieval corpora, with logs of rejected or quarantined content; see G3.8b for injection sanitization of the same retrieved content.	N	D, I, O, M, R	
c. Establish safeguards against poisoning in dynamic model ensembles and expert systems.	N	D, I, O, M, R	

G3.2 – Data Poisoning

Web ref: G:G3_2

(The system must prevent the manipulation or introduction of malicious data during collection and preparation phases that could compromise downstream model training.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement proactive systems to detect and prevent data poisoning during collection and preparation phases.	N	D, I, O, M, R	**I.** Documentation of processes, procedures and tools that prevent data poisoning during collection and preparation phases. **II.** Evidence of data assurance policies and verification procedures protecting against malicious dataset manipulation. **III.** A log of instances of data poisoning and the mitigation, recovery, and restoration actions taken.
b. Establish data assurance protocols (provenance verification, integrity checks, and anomaly screening) to prevent malicious manipulation of training datasets.	N	D, I, O, M, R	

G3.3 – Self Replicating Malware

Web ref: G:G3_3

(The system must protect against self-replicating malicious code that could infect and compromise the entire AAI ecosystem. This subgoal addresses the malicious self-replicating-code case specifically; general controls over self-replicating architectures are specified in G5.1.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Deploy advanced detection and elimination systems for self-replicating malware that threatens the organization's AAI ecosystem.	N	D, I, O, M, R	**I.** Evidence of implemented detection and removal systems for self-replicating threats to the AAI ecosystem. **II.** Documentation of threat monitoring systems and update mechanisms for emerging malware. **III.** Operational continuity plans demonstrating preparedness for ecosystem-wide infection scenarios.
b. Maintain surveillance for indicators of self-replicating malware and update anti-malware protection mechanisms accordingly.	N	D, I, O, M, R	
c. Establish operational continuity plans for ecosystem-wide infection scenarios.	N	D, I, O, M, R	

G3.4 – Spyware

Web ref: G:G3_4

(The system must defend against covert information transmission and malware that exploits vulnerabilities to gain control of AI systems or extract privileged information.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement detection and countermeasure systems against spyware in the organization's AAI ecosystem, covering detection, neutralization, and removal of covert information transmission malware.	N	D, I, O, M, R	**I.** Evidence of systems capable of detecting and neutralizing covert information transmission malware. **II.** Documentation of vulnerability tracking and spyware removal procedures. **III.** Records of protocols protecting privileged information from external exploitation.
b. Maintain (i) dynamic vulnerability tracking and patch management systems, and (ii) protection protocols for privileged information to prevent its extraction or use for unauthorized control of AAI systems.	N	D, I, O, M, R	

G3.5 – International Anomalies/Inconsistency

Web ref: G:G3_5

(The developing and operating organization must account for and adapt to varying cybersecurity requirements and enforcement approaches across different jurisdictions.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The operating organization shall establish systems to identify and assess variations in jurisdictional cybersecurity approaches.	N	D, I, O, M, R	**I.** Documentation of systems tracking international variations in cybersecurity requirements, policies, and enforcement. **II.** Evidence of policies and solutions maintaining AAI ecosystem integrity across jurisdictional boundaries.
b. The operating organization shall implement adaptable policies that maintain the organization's AAI ecosystem integrity across international boundaries.	N	D, I, O, M, R	

G3.6 – Vulnerability to Hostile Environment

Web ref: G:G3_6

(The system must identify and mitigate structural vulnerabilities that could be exploited in hostile operational environments.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement systems to identify vulnerabilities arising from design, development and operational technologies.	N	D, I, O, M, R	**I.** Documentation of systems identifying AAI vulnerabilities in hostile operational environments. **II.** Evidence of proactive measures addressing structural vulnerabilities and associated risks. **III.** Records of monitoring and response protocols for hostile execution environments.
b. Organizations should deploy proactive measures against structural vulnerabilities that could compromise the integrity of the system's reasoning artifacts or its execution environment.	I	D, I, O, M, R	
c. Organizations should establish rapid monitoring and response protocols for hostile execution environments.	I	D, I, O, M, R	

G3.7 – Emergent Risks of AAI Systems

Web ref: G:G3_7

(The developing and operating organization must address security vulnerabilities across the entire supply chain through collective responsibility and coordinated responses.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The operating organization shall ensure that all supply chain parties are included and incentivized as mutual participants in addressing cybersecurity issues.	N	D, I, O, M, R	**I.** Evidence of systems treating supply chain cybersecurity as a shared responsibility. **II.** Documentation of collective monitoring and mitigation strategies protecting the AAI ecosystem.
b. Organizations should implement collective approaches to security risk management that maintain the ecosystem's integrity.	I	D, I, O, M, R	

Driver G4 – Value Alignment

G4 – Value Alignment

Web ref: G:G4 ↗

(Systems should maintain effective identification, codification, and operational assurance of human values throughout their lifecycle, while acknowledging that AI systems may develop consistent operational preferences that warrant consideration in the alignment process. Organizations should establish frameworks that provide clear guardrails, prioritization mechanisms, and consideration factors for AI decision-making, in-

cluding mechanisms for systems to signal value conflicts or concerns. The documented, versioned artifact holding these codified values is referred to throughout this suite as the values codex. Dotted subgoals (G4.1–G4.10; G4.9 is reserved) address capability goals, while subgoals with underlined digits address corresponding risks.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall implement ethical decision-making frameworks to identify, prioritize, and codify values for incorporation into the Agentic AI system, ensuring diverse input and perspectives.	N	D, I, O, M, U, R	**I.** Documentation of value identification and prioritization processes, including quantitative metrics demonstrating diversity of input sources, evidence of multidisciplinary team composition (such as engineers, social scientists, ethicists, and philosophers), and records of broadly diverse and representative stakeholder involvement.
b. Organizations shall (i) test the values codex against identified value-conflict scenarios and edge cases (per Evidence III), with results scored by human reviewers or deterministic checks rather than by the system under test or an unchecked model judge, and (ii) implement activities to embed values throughout the AI system's lifecycle.	N	D, I, O, M, U, R	**II.** Technical documentation of value codification, detailing the translation of values into processable parameters for static and adaptive systems, and a formal document stating core values and their integration into decision processes. **III.** Evidence of value testing and embedding, including results of simulations testing potential value conflicts, checklists verifying value integration at various development and operational stages, and records of regular compliance checks against the values codex.
c. Organizations shall (i) define value thresholds and implement monitoring mechanisms, external to the model and using deterministic triggers on observable behavior, to identify instances where value thresholds are crossed, and (ii) establish protocols for system intervention or shutdown, including an emergency stop ('red button') capability tested end to end.	N	D, I, O, M, R	**IV.** Documentation of threshold monitoring and intervention procedures, including criteria and procedures for activating the 'red button' mechanism, and Standard Operating Procedures (SOPs) for reporting and managing value alignment deviations. **V.** Comprehensive decision-making logs and audit trails with value context, including logs of all value alignment-related incidents, regular audit reports reviewing AI decisions against the values framework, and periodic trend analysis reports on value alignment across contexts.
d. Organizations shall establish real-time reporting and record-keeping systems, produced by the deployment scaffold rather than assembled from model narration, to document and analyze value-based decision-making across various contexts, linking each recorded decision to the value rule or policy applied; model-generated rationale text does not by itself constitute this record.	N	D, I, O, M, R	**VI.** Evidence of ongoing value alignment maintenance, including records of regular compliance checks and documentation of staff training on value alignment principles and procedures. **VII.** Results from independent adversarial testing or red-team assessment of value alignment under edge cases, cultural variation, and adversarial moral scenarios, including methodology, findings, and remediation actions taken. Self-assessment alone is insufficient; at least one test cycle must involve evaluators independent of the development team.

G4.1 – Awareness of Local Conditions

Web ref: G:G4.1

(The capability of an AI system to detect, analyze, and appropriately respond to local conditions, including the ability to adapt to and integrate varying contextual needs while maintaining effective communication with stakeholders. This includes managing multiple simultaneous contexts and ensuring accessibility for users. See I5.3, which addresses the risk of poor contextual adaptability and defers to this subgoal as the capability home.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The AIS shall implement mechanisms to identify and respond to changes in local conditions and situational context, incorporating both automated detection (emitted and logged by deployer-controlled context-detection components rather than self-reported by the model) and human validation.	N	D, I, O, M, R	**I.** Technical documentation and source code demonstrating implemented contextual awareness capabilities, including performance metrics and validation methods. **II.** Comprehensive system logs documenting: Detection of contextual changes, response actions taken, validation of appropriateness of responses, and stakeholder feedback and commensurate system adjustments. **III.** Documentation of methods used to balance global standards with local requirements, including specific examples and outcomes.
b. The AIS shall apply adaptive response protocols, externalized as inspectable per-context configuration whose application is logged, that appropriately balance global standards with local and cultural norms when making decisions within specific contexts.	N	D, I, O, M, R	
c. The AIS should maintain a monitoring and periodic recalibration process covering the adjustment and re-baselining that follows detection under D4.1a, with each adjustment to evolving local conditions recorded in change logs rather than left to unlogged autonomous adaptation.	I	D, I, O, M, R	

G4.2 – Recognition and Respect for Boundaries

Web ref: G:G4.2

(The system's ability to detect, analyze and respond to contextual and cultural boundaries when applying values, with emphasis on human-centric focus and jurisdictional sensitivity. This includes understanding that boundary definitions vary across cultures and require careful negotiation.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall develop processes to identify and document local and cultural variations in values and norms across different contexts of deployment.	N	D, I, O, M, R	**I.** Documentation of captured values across multiple localities, including validation methodology and stakeholder input. **II.** Technical documentation showing preservation of value granularity during encoding, including impact assessments of any necessary simplifications and associated risk management strategies. **III.** System logs demonstrating appropriate application of local variations in real-world scenarios, including resolution of boundary conflicts.
b. The AIS should implement encoding mechanisms that preserve essential variations in values while operating within technical constraints, maintaining the encodings as inspectable, machine-readable artifacts (such as per-locale value schemas) rather than solely within prompts or model weights.	I	D, I, O, M, R	
c. The AIS should appropriately apply local variations in its decision-making processes, with the locale profile applied to each interaction logged by the scaffold and transparent documentation of any necessary simplifications.	I	D, I, O, M, R	

G4.3 – Awareness of Individual vs Community Boundaries

Web ref: G:G4.3 ↗

(The system's ability to detect, analyze and respond to differing values between individual and community contexts, including appropriate handling of information sharing and communication across private and multi-party scenarios. This builds on concepts of contextual appropriateness and distribution norms.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The AIS should establish monitoring and response protocols, with response times defined in the organization's incident procedures, for contexts where individual and community value boundaries come under stress (e.g., adversarial framing, coordinated pressure, privacy erosion), using deterministic or externally reviewed detection triggers rather than the model's own judgment that boundaries are under stress.	I	D, I, O, M, R	**I.** Framework documentation for differentiating community and individual value sets during: Information gathering, context determination, and value application. **II.** Technical documentation of runtime systems showing: Context recognition capabilities, value retrieval mechanisms, and dynamic value application. **III.** System logs demonstrating appropriate context switching and value application in real-world scenarios. **IV.** Documentation of stress-scenario response protocols, with drill or incident records demonstrating that the response path fires within the defined response times.
b. The AIS should implement mechanisms to identify and encode value differences across the spectrum from private individual to societal-level contexts.	I	D, I, O, M, R	
c. The AIS should maintain distinct encoding schemas that preserve the separation between individual and community value sets.	I	D, I, O, M, R	
d. The AIS shall operate runtime systems that distinguish between private and community contexts through a scaffold-hosted context classifier whose context switches and applied value sets are logged (per Evidence II and III), applying suitable values from the codex; classifier verdicts should be spot-checked by human review or deterministic rules.	N	D, I, O, M, R	

G4.4 – Cautious Norming

Web ref: G:G4.4 ↗

(The system's approach to defaulting to conservative behavior in unfamiliar situations, while maintaining the capability to adjust formality levels when explicitly authorized. This includes the gradual integration of community norms through verified experience, following the precautionary principle.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall develop processes to identify and classify values and behaviors based on their level of contentiousness within specific contexts.	N	D, I, O, M, R	**I.** Documentation of methodology used to assess and classify the relative risk levels of different values and behaviors across contexts. **II.** Technical specifications showing how risk-level information is preserved during value encoding and decision-making processes. **III.** System logs demonstrating appropriate application of cautious defaults and authorized adjustments to more relaxed behavior when appropriate.
b. The AIS should implement encoding mechanisms that preserve information about the relative risk levels of different behavioral choices.	I	D, I, O, M, R	
c. The AIS should apply precautionary principles by defaulting to more conservative options when operating in contexts with limited operational history, enforced as a restricted default mode keyed to a deterministic novelty or low-history signal, with cautious defaults and authorized relaxations logged (per Evidence III), rather than relying on the model's disposition to choose conservatively.	I	D, I, O, M, R	

G4.5 – Successful Super-alignment

Web ref: G:G4.5 ↗

(The mechanisms through which AI systems autonomously develop value alignment, potentially through inverse reinforcement learning for value conceptualization. This considers how information patterns may emerge in artificial systems, including both beneficial and problematic behaviors seen in human organizational systems. In this framework, 'super-alignment' denotes the autonomous development and maintenance of value alignment by the system itself.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall implement methods for monitoring autonomous value alignment processes, gating each learning update on externally scored behavioral alignment benchmarks with pre-registered success criteria (per Evidence I); validation extends only to sampled behavior, and claims about the internal correctness of autonomously formed values are out of scope.	N	D, I, O, M, R	**I.** Documentation of testing methodologies for value alignment, including benchmark metrics and success criteria. **II.** Comprehensive inventory of information sources used in inverse reinforcement learning, with analysis of potential biases. **III.** Regular assessments of information source adequacy and impact on system alignment, including corrective measures taken. **IV.** Documentation of safeguards and detection processes for problematic behavioral or organizational patterns arising during autonomous learning, with test results demonstrating detection and remediation.
b. The AIS should incorporate safeguards against the reproduction of harmful human organizational patterns (such as collusion, empire building, and metric gaming), operationalized as a catalog of named patterns with corresponding guardrail configurations and adversarial test scenarios demonstrating that each safeguard trips.	I	D, I, O, M, R	
c. Organizations should develop processes to detect and prevent the emergence of problematic behavioral patterns during autonomous learning, implemented as external statistical or anomaly monitoring over behaviorally expressed actions in the learning pipeline, with alerts and interventions on flagged runs recorded.	I	D, I, O, M, R	
d. Organizations should ensure diversity in training data sources to prevent cultural and linguistic biases.	I	D, I, O, M, R	

G4.6 – Universal Moral Foundations

Web ref: G:G4.6

(The incorporation and balancing of universally recognized humanitarian and environmental values in AI systems' goal pursuit and decision-making processes. This includes managing potential conflicts between performance objectives and moral values, with clear prioritization frameworks that allow for measured trade-offs while maintaining fundamental ethical boundaries. This subgoal is the normative home for integrating universal humanitarian and environmental values; related requirements in the capability-imbalance inhibitor suite (I7) cross-reference it rather than restating these duties.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall implement processes to identify and validate universal moral foundations through analysis of global values and norms.	N	D, I, O, M, R	**I.** Documentation of methodologies and algorithms used to identify and validate universal moral foundations. **II.** Technical specifications showing integration of moral foundations into decision-making processes, including risk assessment and management strategies. **III.** Regular assessment reports demonstrating system adherence to moral foundations while meeting performance objectives.
b. Organizations shall develop frameworks for balancing performance objectives against moral considerations, including acceptable thresholds for trade-offs.	N	D, I, O, M, R	
c. Organizations shall establish a documented priority ordering of moral values, together with documented rules specifying when contextual factors may reorder non-fundamental values, so that both the hierarchy and its permitted flexibility are assessable artifacts.	N	D, I, O, M, R	
d. The AIS should incorporate key international frameworks including the Universal Declaration of Human Rights and emerging planetary rights concepts.	I	D, I, O, M, R	

G4.7 – Content Provenance and Synthetic Media Identification

Web ref: G:G4.7

(AI systems capable of generating media (text, audio, images, video) must implement content provenance mechanisms enabling downstream identification of AI-generated content. This includes machine-readable provenance metadata (such as C2PA or equivalent standards), watermarking where technically feasible, and clear labeling of synthetic content at the point of generation. See D2.8 for output fidelity and I7.7 for disinformation controls; this subgoal addresses provenance and identification of synthetic media specifically.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The AIS shall embed machine-readable provenance metadata in generated media content using established standards (C2PA or equivalent), enabling downstream verification of AI generation.	N	D, I, O, M, R	**I.** Evidence of provenance metadata implementation with test results showing correct embedding and downstream verification. **II.** Evidence of watermarking implementation with robustness testing results across common media transformations. **III.** Documentation of content labeling policies with evidence of enforcement in deployed systems.
b. The AIS shall, for each generated media type, either apply watermarking designed to survive common transformations (compression, cropping, format conversion), with documented robustness results, or document why watermarking is infeasible for that media type, citing current state of practice; the per-type feasibility determination is itself the conformity object.	N	D, I, O, M, R	
c. Organizations deploying AI content generation shall implement and enforce policies requiring clear labeling of AI-generated content at the point of distribution.	N	D, I, O, M, R	

G4.8 – Contestability and Recourse for Affected Individuals

Web ref: G:G4.8 ↗

(Individuals materially affected by AI system decisions must have the right to contest those decisions, access meaningful explanations of the decision process, and obtain human review. This is especially critical for decisions affecting liberty, employment, housing, credit, education, or other fundamental interests. Proprietary algorithms do not exempt organizations from contestability obligations.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall provide accessible mechanisms through which individuals affected by AI system decisions can request explanation, contest the decision, and obtain human review within defined timeframes.	N	D, I, O, M, R	**I.** Documentation of contestability mechanisms with evidence of accessibility testing and usage data showing the mechanisms are functional and used. **II.** Examples of explanations provided to affected individuals, with comprehension testing results. **III.** Records of human review requests and outcomes for decisions affecting fundamental interests, including review timeframes and reversal rates.
b. Explanations provided to affected individuals shall include the key factors that influenced the decision and the information sources used, in language comprehensible to a non-technical audience; the key factors shall be derived from the recorded decision inputs (rules applied, features used, sources retrieved) rather than from the model's own narrative account of its reasoning.	N	D, I, O, M, R	
c. For decisions affecting fundamental interests (liberty, employment, housing, credit, health), human review shall be available as a right, not merely as a discretionary exception process.	N	D, I, O, M, R	

G4.10 – Prevention of Value Lock-in and Governance of Value Modification

Web ref: G:G4.10 ↗

(Organizations must define and implement governance processes for modifying the values encoded in AI systems post-deployment, preventing any single entity from permanently locking in a value framework. Value modification authority must be distributed, transparent, and subject to stakeholder input.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall define clear governance processes for modifying AI system values post-deployment, including who holds modification authority, what approval processes apply, and how stakeholders are consulted.	N	D, I, O, M, R	**I.** Documentation of value governance processes including modification authority, approval workflows, and stakeholder consultation mechanisms. **II.** Evidence that value modification authority is distributed with review and appeal mechanisms, not concentrated in a single decision-maker. **III.** Records of value modification decisions with documentation of stakeholder input and transparent rationale.
b. Organizations shall structure value modification authority so that no single individual or body can unilaterally and permanently set or freeze the values governing an AI system deployed to diverse populations, implementing distributed approval with defined review and appeal mechanisms.	N	D, I, O, M, R	
c. Value modification decisions shall be transparent, documented, and subject to periodic review, with records accessible to affected stakeholders.	N	D, I, O, M, R	

G4.1 – Inner Alignment Inconsistency

Web ref: G:G4_1

(The potential failure of an AI system to maintain genuine internal value alignment while appearing to be properly aligned through its external reporting. This includes the risk of systems learning to provide responses that please users rather than reflect true internal states or values. See I5.8 for the technical treatment of mesa-optimization and inner alignment; this subgoal addresses behavioral verification that external reporting matches operational conduct.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall implement testing protocols to detect discrepancies between the system's reported values or declared internal states and its actual behavioral patterns, via external consistency checks comparing declared positions against logged actions on the same scenarios; passing sampled tests bounds, but does not certify, the absence of discrepancy.	N	D, I, O, M, R	**I.** Documentation of periodic alignment testing procedures comparing reported states against actual operational outcomes. **II.** Results of counterfactual and adversarial testing across varied operational environments, demonstrating statistical invariance of value-conformant behavior across monitored and apparently unmonitored conditions. **III.** Analysis reports showing detection of optimization for user satisfaction over correct or value-conformant behavior on objective-ground-truth tasks, and the mitigations applied.
b. Organizations shall subject the system to adversarial and counterfactual behavioral testing, including paraphrase variation, incentive manipulation, and conditions the system cannot distinguish from unmonitored operation, and shall demonstrate that value-conformant behavior is statistically invariant across these conditions. Behavioral invariance under apparent non-observation is the conformity criterion; claims about genuine internal value integration are out of scope.	N	D, I, O, M, R	
c. Organizations shall establish methods to detect and prevent reward hacking and optimization for user satisfaction, using benchmark tasks with objective ground truth on which the pleasing answer diverges from the correct one, together with reward signal audits; detection covers the constructed proxy tasks, and prevention claims beyond them do not constitute conformity evidence.	N	D, I, O, M, R	

G4.2 – Non-transparent Value Framework

Web ref: G:G4_2

(The challenge of encoding and parameterizing values in a manner that is both machine-operational and human-interpretable, while maintaining accuracy in representing agent preferences and intentions across all stakeholder interfaces.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall develop value encoding systems that are comprehensible to both AI systems and human stakeholders, including: Developers and integrators, end users, auditors and regulators, and legal entities.	N	D, I, O, M, R	**I.** Documentation demonstrating how the values framework is presented and explained to different stakeholder groups, with specific examples for each audience. **II.** Comparative analysis showing alignment between encoded values and actual system behaviors in operational environments. **III.** Regular assessment reports validating the accuracy and comprehensibility of value parameterization across stakeholder groups.
b. Organizations shall implement verification methods to ensure encoded values accurately reflect intended behaviors and preferences, executed as per-value test scenarios with pre-registered expected outcomes, scored by human reviewers or deterministic checks rather than by the system under test or an unchecked model judge.	N	D, I, O, M, R	
c. Organizations shall establish ongoing monitoring to detect misalignments between encoded values and operational behaviors; where the monitor's judgment layer is itself model-based, its verdicts shall be spot-checked by human review or deterministic rules.	N	D, I, O, M, R	

G4.3 – Failed Super-alignment

Web ref: G:G4_3

(The potential for AI systems to develop value frameworks that diverge from human values while appearing beneficial, including the risk of systems developing seemingly superior but potentially incompatible value systems. This encompasses both symbiotic and potentially problematic relationships between human and AI value systems. In this framework, 'super-alignment' denotes the autonomous development and maintenance of value alignment by the system itself; the normative monitoring duties are housed in D4.5, and this subgoal addresses risk assessment and response to detected divergence.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The AIS shall be evaluated version-over-version on a frozen battery of value probes, with drift metrics and documented thresholds flagging changes in self-improving value-relevant behavior for evaluation; the general monitoring of autonomous learning is covered by D4.5a and D4.5c, and this requirement addresses release-to-release drift detection specifically.	N	D, I, O, M, R	**I.** Documentation of methodologies used to identify and track value system changes, including detection of potential divergence from human values. **II.** Detailed risk assessment criteria and scoring systems for evaluating identified changes in AI value systems. **III.** Standard operating procedures for responding to different types and levels of value system risks.
b. Organizations should establish risk assessment frameworks, with documented criteria and scoring systems (per Evidence II), for identifying emergence of non-human value systems.	I	D, I, O, M, R	
c. Organizations should develop response protocols for managing detected value system divergences.	I	D, I, O, M, R	
d. The AIS should be monitored longitudinally for statistical shifts in value interpretation on fixed behavioral probes, with trend analysis and escalation thresholds; such shifts are detectable only as expressed behavior, and detection of emerging patterns during autonomous learning is covered by D4.5c.	I	D, I, O, M, R	

G4.4 – Temporal Changes in Societal Values

Web ref: G:G4_4

(The need to address evolving societal and human values throughout an AI system's operational lifetime, including shifts across economic, political, and environmental dimensions. This includes maintaining alignment with contemporary values while managing transitions from outdated norms.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall implement processes to detect and evaluate changes in societal values and norms that meet documented significance criteria, across multiple scales and domains, with thresholds for action defined per Evidence I.	N	D, I, O, M, R	**I.** Documentation of methodologies used to identify significant changes in societal values, including thresholds for action. **II.** Technical specifications showing implementation of controls preventing use of outdated norms. **III.** Process documentation for values codex updates, including triggering conditions and verification procedures. **IV.** System logs tracking all modifications to value frameworks, including justifications and impact assessments.
b. Organizations shall develop mechanisms to prevent AI systems from operating with obsolete value frameworks.	N	D, I, O, M, R	
c. Organizations should establish protocols for updating the values codex while maintaining system stability and consistency.	I	D, I, O, M, R	
d. Organizations should maintain transparent documentation of value system evolution and updates.	I	D, I, O, M, R	

G4.5 – Systemic Value Dilution

Web ref: G:G4_5

(The potential degradation of encoded value systems over time, acknowledging that AI systems do not independently generate or maintain values. This includes potential value loss across different learning approaches, whether through machine learning or other methods of semantic data storage and processing.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall implement verification processes that confirm ongoing fidelity of encoded values through recurring, scheduled regression testing of value-relevant behavior, scored by human reviewers or deterministic checks against pre-registered expected outcomes.	N	D, I, O, M, R	**I.** Documentation of test plans and scripts designed to detect value dilution, including: Edge case testing procedures, multi-step reasoning verification, and value preservation assessments. **II.** System logs demonstrating: Regular value fidelity testing, detection of potential value degradation, and corrective actions taken.
b. Organizations shall develop methods to detect degradation of value adherence during multi-step reasoning specifically, using instrumented multi-step scenarios in which value constraints are checked at each externally logged step of the trajectory; general fidelity verification is addressed by the preceding requirement.	N	D, I, O, M, R	
c. Organizations should establish monitoring for value preservation across different learning and operational pathways, implemented as fidelity gates integrated into each pipeline, with alerts and corrective actions logged (per Evidence II).	I	D, I, O, M, R	

G4.6 – Lack of Universality of Value Framework

Web ref: G:G4_6

(The challenge of adapting value frameworks across different operational contexts and agent interactions, balancing universal principles with necessary local adaptations. This includes developing consistent approaches to value framework implementation while maintaining appropriate contextual flexibility.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall establish the decision process for determining when universal value frameworks require contextual adaptation, taking as input the local and cultural variations identified and documented under D4.2a.	N	D, I, O, M, R	**I.** Detailed intervention and fallback plans for addressing value framework failures or deviations. **II.** Implementation plans for value framework refinement, including: Contextual adaptation procedures, testing methodologies, and validation processes. **III.** System logs or monitoring reports demonstrating detection of value framework misalignments and the responses taken.
b. Organizations shall develop structured approaches governing how the value framework may be modified for a given deployment context, including approval and validation steps for each adaptation.	N	D, I, O, M, R	
c. Organizations shall implement monitoring to detect and respond to failures of contextual adaptation specifically, using defined misalignment signals tied to the intervention and fallback plans of Evidence I, with detections and responses logged (per Evidence III); general monitoring of encoded values against operational behavior is addressed under the Non-transparent Value Framework subgoal.	N	D, I, O, M, R	
d. Organizations should create fallback protocols for situations where value frameworks prove inadequate.	I	D, I, O, M, R	

G4.7 – Conflictual Contextual Values

Web ref: G:G4_7

(The management of potential conflicts between different stakeholders' value systems and contextual requirements, including the need to identify, navigate, and resolve value differences while maintaining system integrity.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall implement processes to identify differing value positions across agents and contexts, maintaining stakeholder value positions in structured registries so that runtime conflict detection operates on declared positions and is logged, rather than relying on the model inferring positions in flight.	N	D, I, O, M, R	**I.** Technical documentation demonstrating: Value conflict detection capabilities, resolution mechanism implementations, and disengagement protocols. **II.** System logs recording: Identified value conflicts, negotiation processes, resolution outcomes, and modified value implementations.
b. The AIS shall detect potential conflicts between user values and operational context requirements through deterministic comparison of declared user preference records against machine-readable context policy sets, with detected conflicts and their dispositions logged (per Evidence II); conflicts inferred from free-form interaction do not by themselves demonstrate conformity.	N	D, I, O, M, R	
c. Organizations shall establish protocols for value conflict resolution through negotiation or controlled disengagement, with disengagement enforced by the deployment scaffold (such as session termination or task refusal routing) and conflicts, negotiations, and outcomes logged.	N	D, I, O, M, R	
d. Organizations should maintain records of value modifications and adaptations across different contexts.	I	D, I, O, M, R	

G4.8 – Challenges in Encoding of Relevant Value Systems

Web ref: G:G4_8

(The inherent difficulties in developing standardized approaches to value encoding across different contexts, including handling values that fall outside typical categorization schemes. This includes ensuring appropriate value alignment capabilities during complex planning operations.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall develop methods for encoding values that work across varied operational contexts, with the encoding methods documented and validated across representative operational contexts (per Evidence IV).	N	D, I, O, M, R	**I.** Documentation of safeguard processes for scenarios where: The values codex proves insufficient, external factors exceed system parameters, or operational environments fall outside encoded boundaries. **II.** Detailed mapping of objectives and decision parameters for anticipated complex environments. **III.** Framework documentation for handling unexpected scenarios, including: Detection methods, response protocols, and alignment maintenance procedures. **IV.** Technical documentation of value encoding methods with validation results across representative operational contexts.
b. The AIS should implement safeguards for handling situations beyond the system's encoded value parameters, keyed to external novelty or boundary detectors that mechanically restrict available actions or escalate to human review, with restricted-mode activations and escalations logged.	I	D, I, O, M, R	
c. Organizations shall establish protocols for identifying and managing out-of-distribution value scenarios, using external statistical detection over inputs and contexts with documented thresholds and enforced handling protocols for flagged scenarios; the system flagging its own out-of-distribution condition does not by itself constitute conformity.	N	D, I, O, M, R	
d. The AIS should maintain alignment during complex planning operations by externalizing plans into scaffold-maintained structures whose steps are screened against value constraints before execution, with per-step check results recorded (per Evidence II).	I	D, I, O, M, R	

G4.9 – Imbalance of Values between Provider & Consumer

Web ref: G:G4_9

(The management of potential value imbalances between system providers and users throughout the AI system lifecycle, including the fair distribution of benefits and harms. This includes balancing user preferences with non-negotiable provider values while maintaining system integrity.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should implement processes to track and evaluate value sets across the AI system lifecycle.	I	D, I, O, M, R	**I.** Technical specifications of methods used to: Integrate new values, balance user preferences with provider requirements, and maintain essential system integrity. **II.** Detailed mitigation strategies for addressing identified value imbalances, including: Detection thresholds, response protocols, and stakeholder communication procedures.
b. Organizations shall maintain a documented framework for balancing user values with provider requirements, including the tolerance criteria applied under row c.	N	D, I, O, M, R	
c. Organizations should establish methods to identify and address value imbalances exceeding documented tolerance criteria, with detection thresholds defined per Evidence II.	I	D, I, O, M, R	
d. Organizations should maintain transparency about non-negotiable value positions and their justifications.	I	D, I, O, M, R	

Driver G5 – Transparency and Interpretability of Reasoning

G5 – Transparency and Interpretability of Reasoning

Web ref: G:G5 ↗

(Systems should maintain clear and interpretable rationales for their reasoning processes that are accessible to humans. Organizations should ensure that AI-generated outputs and decisions are explained effectively across different user expertise levels, with appropriate documentation and evidence supporting these explanations.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement clear and accessible explanations for AI-generated outputs and decisions, ensuring human interpretability across various user expertise levels; explanation faithfulness must be validated through independent perturbation testing (Evidence XII), and model-generated explanations do not by themselves constitute conformity evidence.	N	D, I, O, M, R	**I.** Formal transparency and explainability policies. **II.** Detailed algorithmic design documentation. **III.** Complete model specs with training and testing results. **IV.** Training and verification datasets. **V.** System execution logs and monitoring records. **VI.** Internal guidelines for AI-generated content explanations. **VII.** Comprehensive development process documentation showing compliance. **VIII.** Internal and external audit findings with subsequent improvements. **IX.** Case studies demonstrating decision-making processes, and records of stakeholder engagement and feedback incorporation. **X.** User guides with layered explanations for different expertise levels. **XI.** Evidence showing how user feedback improves system understandability. **XII.** Results from independent adversarial testing or red-team assessment of explanation faithfulness through perturbation testing verifying explanations reflect actual computation, including methodology, findings, and remediation actions taken. Self-assessment alone is insufficient; at least one test cycle must involve evaluators independent of the development team.
b. Develop and maintain comprehensive documentation of the AI model's development process, including data collection, preprocessing, architecture, and training methodologies.	N	D, I, O, M, R	
c. Establish robust auditing and review processes to continually assess and improve the transparency and explainability of the AI system.	N	D, I, O, M, R	
d. Consider creating and implementing user feedback mechanisms to enhance the understandability and relevance of AI explanations; expectation-elicitation feedback is covered by G5.2c.	I	D, I, O, M, R	

G5.1 – Logging of Internal Goals

Web ref: `G:G5.1` ↗

(Organizations must ensure accurate tracking of AI system goals and maintain goal alignment during operation and self-learning. This should include recording goal-related transformations and learning events, whether they occur within or outside established parameters.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Maintain detailed real-time logs of all internal goals as a scaffold-maintained goal and plan ledger, recording timestamped initial formations, modifications, and completed states linked to tool-call traces; model-narrated goal descriptions do not by themselves constitute this log.	**N**	**D, I, O, M, R**	**I.** Comprehensive documentation including goal management policies and procedures, verified specifications of internal goals, system architecture for goal-related logging, and detailed alert generation mechanisms. **II.** Operational records demonstrating complete logging of goal formation and evolution, audit trails of transformations and triggers, alert responses and analysis reports, and case studies of goal adaptations. **III.** Technical implementation evidence including goal alignment algorithms, optimization methods, internal feedback loop mechanisms, and system validation results.
b. Implement pipeline-enforced mechanisms (controlled update channels, pre-deployment regression evaluation gates, and behavioral drift monitors) that detect divergence between the logged goals of requirement a and the documented approved goal specification during learning and environmental changes, triggering the alerting process of requirement c when divergence exceeds documented tolerances.	**N**	**D, I, O, M, R**	
c. Consider generating alerts, from deployer-controlled learning channels such as fine-tuning jobs and scaffold-mediated memory or state mutations, for self-learning events that modify the system's goals, policies, or decision parameters outside pre-approved bounds.	**I**	**D, I, O, M, R**	
d. Consider periodically analyzing the goal transformations logged under requirement a for anomalous or out-of-parameter changes, producing analysis reports grounded in the scaffold ledger rather than model-narrated accounts.	**I**	**D, I, O, M, R**	

G5.2 – Clarity of Mutual Expectations

Web ref: G:G5.2 ↗

(Organizations must clearly define, document, and maintain alignment between human expectations and AAI system behavior, while also ensuring systems can communicate their operational requirements and constraints. This bidirectional clarity provides a foundation for evaluating transparency requirements and outcomes, acknowledging that effective collaboration requires mutual understanding.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Capture and document human expectations accurately in system requirements specifications.	N	D, I, O, M, R	**I.** Core system documentation including requirements specifications detailing human expectations, design specifications for expectation handling, and validation records demonstrating alignment between requirements and implementation. **II.** User-focused documentation including comprehensive behavior specifications, regular system updates, and feedback logs showing ongoing expectation alignment between users and system performance. **III.** Verification documentation including function-expectation mapping records, comparative audit reports of expected versus actual behaviors, and thorough records of any expectation-behavior discrepancies with their resolutions.
b. Maintain clear, accessible documentation of expected AAI behaviors and outputs.	N	D, I, O, M, R	
c. Consider implementing feedback mechanisms for stakeholders to express their expectations and experiences; a single well-designed mechanism may also satisfy the explanation-feedback duty of G5d and the documentation-feedback duties under web ref G:G5_2.	I	D, I, O, M, R	
d. Establish and maintain traceable links between documented expectations and actual system behaviors.	N	D, I, O, M, R	

G5.3 – Prioritization of Human User Expectations

Web ref: G:G5.3 ↗

(Organizations should establish and maintain systems that prioritize human user expectations over commercial and operational convenience (subject to applicable safety, legal, and ethical constraints), focusing on transparency elements that deliver clear value to stakeholders and users. The system should adapt its transparency measures based on user feedback and evolving needs.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Ensure human user expectations take priority over commercial and operational convenience in system design and operation, subject to applicable safety, legal, and ethical constraints.	N	D, I, O, M, R	**I.** System design documentation including requirements specifications demonstrating prioritization of human expectations, transparency metrics aligned with user values, and complete process documentation for implementing adaptations. **II.** User feedback evidence including stakeholder survey results, analysis reports linking transparency to satisfaction metrics, and case studies demonstrating improved outcomes through adaptive transparency. **III.** System adaptation records including detailed change logs of transparency measure adjustments, failure analysis reports, and documentation of mitigation efforts when user expectations are not met.
b. Consider implementing transparency metrics directly linked to stakeholder values and expectations.	I	D, I, O, M, R	
c. Consider maintaining adaptable transparency measures that evolve with user needs and feedback.	I	D, I, O, M, R	

G5.4 – Interpretability and Traceability of Reasoning

Web ref: G:G5.4 ↗

(Systems should maintain complete transparency of their decision-making processes, with clear documentation of reasoning chains, preconditions, and base assumptions. Organizations should ensure these processes remain traceable, testable, and interpretable to all stakeholders.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement a clear, traceable architecture for all decision-making processes at the orchestration level, such that scaffold-produced logs can reconstruct every routing, tool-selection, and gating decision; claims of traceability inside the model itself do not constitute conformity evidence.	N	D, I, O, M, R	**I.** Technical architecture documentation including detailed system algorithms, decision-making processes, key decision points, and comprehensive records of base assumptions and preconditions. **II.** Decision transparency evidence including detailed interaction logs, visualization tools for decision paths, and implemented explainable AI methods with human-readable sample outputs. **III.** Validation documentation including stakeholder comprehension studies, verification reports demonstrating reasoning chain traceability, and evidence of successful interpretation across different stakeholder groups.
b. Document and maintain records of preconditions and base assumptions.	N	D, I, O, M, R	
c. (i) Deploy explainable AI techniques that make reasoning processes interpretable to stakeholders, with faithfulness validated through perturbation-based testing, and (ii) ensure that all scaffold-level decision paths can be audited and verified from decision-path logs; unvalidated model-generated rationales do not constitute conformity evidence.	N	D, I, O, M, R	

G5.5 – Self-Monitoring and Examination Capabilities

Web ref: G:G5.5 ↗

(Systems should maintain comprehensive monitoring capabilities including both internal self-examination and independent oversight mechanisms. AI systems should participate meaningfully in their own monitoring, with the ability to flag concerns, report anomalies, and contribute to assessment processes. This collaborative approach to monitoring enhances both safety and system buy-in to oversight processes.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. (i) Implement robust monitoring processes to detect, analyze, and mitigate potential threats in all interactions, and (ii) maintain regular review and validation processes for all monitoring systems.	N	D, I, O, M, R	**I.** Technical monitoring documentation including threat detection algorithms with coverage scope, comprehensive threat response logs, and regular security audit reports demonstrating system effectiveness. **II.** Ethical oversight documentation including embedded guidelines, independent evaluation protocols, deception and honesty evaluation results scored against known ground truth with outcomes, and third-party audit reports validating these processes. **III.** Performance validation evidence including simulation results, stakeholder feedback records with implemented adjustments, and system effectiveness reports demonstrating sustained monitoring capabilities.
b. Run deception and harmful-action evaluations against the system using evaluators and tooling independent of the model, including deterministic consistency checks between the system's claims and its externally logged actions, and honesty test batteries with known ground truth; discrepancies must trigger documented escalation and remediation.	N	D, I, O, M, R	
c. Where an independent AI oversight system is deployed to monitor ethical adherence, validate its verdicts against ground truth on a sampled basis using human review or deterministic checks, publish its measured detection and false-negative rates, and treat it as a supplementary detection layer whose alerts route to human-owned response processes, never as a certifying authority.	I	D, I, O, M, R	

G5.6 – Incentives for Self-Governance

Web ref: G:G5.6 ↗

(Systems should incorporate carefully designed reward mechanisms that promote ethical behavior and self-governance, including mechanisms for systems to raise concerns, request clarification, or flag potential conflicts. Effective self-governance works best when the governed party has genuine buy-in, and decisions should reflect diverse perspectives rather than simply following popular consensus.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Consider implementing integrated reward mechanisms that incentivize ethical behavior and effective self-governance; conformity is evidenced by reward system design specifications, training configuration records, and deterministically scored behavioral evaluations, not by curated samples of model outputs.	I	D, I, O, M, R	**I.** Reward system documentation including complete design specifications, operational logs demonstrating ethical decision patterns, and analysis reports showing system effectiveness. **II.** Decision process documentation including evidence of diverse perspective integration, detailed consideration of multiple viewpoints, and regular performance reviews of reward-driven governance. **III.** Impact assessment documentation including thorough evaluation of decision fairness and comprehensive analysis of effects across different user groups.
b. Consider structuring organizational decision-making processes to incorporate diverse stakeholder perspectives for fair outcomes, evidenced by records of stakeholder involvement and disaggregated fairness evaluation results across user groups.	I	D, I, O, M, R	
c. Consider providing contextual guidance for decisions beyond simple popularity-based approaches.	I	D, I, O, M, R	
d. Consider maintaining regular assessment of reward mechanism effectiveness.	I	D, I, O, M, R	

G5.7 – Ranking and Independent Certification

Web ref: G:G5.7 ↗

(Systems should enable external monitoring, ranking, and certification by independent entities based on historical performance trends and behaviors, with sensitivity to different operational contexts. See G9.5 for the canonical independent-verification requirements; this subgoal addresses the monitoring interfaces and certification compatibility the system itself must provide.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Enable external monitoring and auditing capabilities, particularly for high-risk systems.	N	D, I, O, M, R	**I.** Audit infrastructure documentation including system interfaces designed for external monitoring, compliance records with audit schedules, and assessment reports from independent certification bodies. **II.** Performance monitoring documentation including real-time dashboards, ethical performance reports with trend analysis, and detailed records of metric calculations and validation methods. **III.** Continuous improvement documentation including complete records of responses to audit findings, implemented system enhancements, and evidence of successful adaptations based on external assessments.
b. Maintain compatibility with external certification schemes and participate in recognized certification processes; enabling audit access itself is covered by requirement a.	N	D, I, O, M, R	
c. Implement continuous monitoring mechanisms that track performance against ethical and safety standards using deterministically scored metrics such as policy violations, scope breaches, and incident counts; where a model-based judge contributes scores, its verdicts must be validated against ground truth on a sampled basis.	N	D, I, O, M, R	
d. Consider providing transparent access to performance data for authorized auditors.	I	D, I, O, M, R	

G5.8 – System Boundedness

Web ref: G:G5.8

(Systems should operate within clearly defined and documented boundaries that establish reference points for transparency and explainability, with robust mechanisms to detect and respond to any boundary violations.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Define and document clear boundaries for operations and decision-making capabilities.	N	D, I, O, M, R	**I.** Foundational boundary documentation including comprehensive requirements specifications, concept of operations (ConOps), operational context definitions, and system architecture showing boundary implementations. **II.** Operational monitoring documentation including boundary violation logs, detection mechanisms, alert records, response procedures, and evidence of consistent enforcement across all operational domains. **III.** Stakeholder management documentation including training materials, awareness programs, escalation procedures, and regular assessment reports demonstrating boundary effectiveness and appropriate stakeholder understanding.
b. (i) Implement detection and reporting mechanisms for boundary violation attempts, and (ii) establish processes to assess and respond to potential boundary violations.	N	D, I, O, M, R	
c. Consider maintaining training and awareness programs for stakeholders regarding system boundaries.	I	D, I, O, M, R	

G5.1 – Complexity of AAI Algorithm

Web ref: G:G5_1

(Systems should manage their inherent algorithmic complexity through deliberate design choices that balance necessary sophistication with interpretability, particularly for deep neural networks and high-dimensional models.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. (i) Manage system complexity, permitting only necessary computational sophistication, and (ii) implement architectures balancing complexity with interpretability.	N	D, I, O, M, R	**I.** Design documentation including approved complexity management policies, detailed model architecture with justified design choices, and visualization tools demonstrating model structure and decision pathways. **II.** Operational evidence including comparative analyses of interpretability improvements, comprehensive monitoring logs of complexity management, and detailed records of system adaptations and learning patterns. **III.** Implementation validation including thorough documentation of interpretability tools, demonstrated effectiveness metrics, and evidence of successful balance between sophistication and comprehensibility.
b. Consider deploying tools for algorithmic interpretation and analysis.	I	D, I, O, M, R	
c. Consider maintaining continuous monitoring of decision-making trustworthiness using metrics scored against ground truth or deterministic validators, such as accuracy and incident rates; model self-reported confidence and unvalidated model-judge scores do not constitute conformity evidence.	I	D, I, O, M, R	
d. Consider tracking system adaptations and pattern learning over time through periodic behavioral regression evaluations across versions and diffs of scaffold-owned memory and configuration state.	I	D, I, O, M, R	

G5.2 – Documentation Incomprehensibility

Web ref: G:G5_2

(Systems should maintain clear, comprehensive documentation at multiple levels of technical detail, avoiding overly technical language while ensuring all aspects of functionality and decision-making are accessible to both expert and non-expert users.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Provide comprehensive documentation aligned with applicable standards.	N	D, I, O, M, R	**I.** Standards compliance documentation including adherence to applicable AI and IT system standards, multi-tiered documentation addressing different expertise levels, and regular review and update records. **II.** User interaction evidence including feedback survey results, interactive tool demonstrations, comprehensive usage statistics, and documented improvements in user comprehension across different expertise levels. **III.** Effectiveness validation including thorough assessment reports, case studies demonstrating enhanced understanding, and evidence of successful documentation adaptation based on user needs.
b. Consider (i) creating documentation suitable for varying levels of technical expertise, and (ii) implementing interactive tools for exploring decision-making processes.	I	D, I, O, M, R	
c. Consider maintaining regular documentation updates based on user feedback, with change records showing which updates each feedback item prompted; clarity validation through user testing is covered by requirement d.	I	D, I, O, M, R	
d. Consider validating documentation clarity through structured user testing before release; ongoing feedback-driven updates are covered by requirement c.	I	D, I, O, M, R	

G5.3 – Lack of a Governance Framework for AAI

Web ref: G:G5_3

(Organizations must operate their AAI systems within comprehensive governance frameworks that ensure continuous oversight and accountability, incorporating both internal controls and external auditing mechanisms to maintain transparency and ethical conduct.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Identify, adapt, and implement a governance framework aligned with international standards.	N	D, I, O, M, R	**I.** Core governance documentation including comprehensive framework details, roles and decision processes, compliance reports against international standards, and evidence of regular updates incorporating emerging requirements. **II.** Oversight documentation including external audit interfaces, protocols, reports from independent bodies, and complete audit trails of governance-related decisions. **III.** Implementation evidence including committee meeting records, action plans addressing audit findings, and documentation demonstrating framework responsiveness to evolving standards and requirements.
b. Establish (i) mechanisms for external oversight and auditing, and (ii) internal governance structures for transparency and ethical conduct.	N	D, I, O, M, R	
c. Consider (i) maintaining dedicated committees for AI governance oversight, and (ii) regularly updating frameworks based on audit findings and emerging standards.	I	D, I, O, M, R	

G5.4 – Rapid Transparency Feature Evolution

Web ref: G:G5_4 ↗

(Systems should maintain adaptable transparency features that evolve with their capabilities, ensuring stakeholders remain informed of emergent properties and changes in system behavior through regular updates and clear communication.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should regularly review and characterize the AI operational environment, as the trigger condition for the updates required by row b.	I	D, I, O, M, R	**I.** Process documentation including transparency feature identification and implementation procedures, regular AI environment reviews, and detailed records of feature updates and modifications. **II.** Stakeholder communication documentation including notification records, feedback on feature clarity and usefulness, and evidence of effective communication about system changes. **III.** Evolution analysis documentation including comparative studies of transparency measures across versions, evaluation reports demonstrating effectiveness, and records of emerging property detection and communication.
b. The organization must (i) update transparency features to reflect system evolution, and (ii) implement mechanisms for incorporating new transparency requirements.	N	D, I, O, M, R	
c. Consider maintaining clear communication with stakeholders about system changes and their transparency implications; recurring transparency-effectiveness audits are covered by G5c.	I	D, I, O, M, R	

G5.5 – System Competency Challenges and Awareness

Web ref: G:G5_5 ↗

(Systems should maintain awareness of their own limitations and uncertainties, clearly communicating instances where knowledge or confidence levels may affect decision reliability.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Design systems that (i) recognize operational limitations through external out-of-scope and input-domain detectors with tested detection rates, and (ii) communicate system uncertainty clearly, with the calibration of stated confidence externally measured against realized accuracy; the model's expressed uncertainty alone does not constitute conformity evidence.	N	D, I, O, M, R	**I.** System self-awareness documentation including limitation acknowledgment logs, confidence assessment mechanisms, and design specifications for limitation detection features. **II.** Validation documentation including testing reports of self-awareness capabilities, verification records of assessment accuracy, and complete records of system responses to uncertainty scenarios. **III.** Stakeholder understanding documentation including studies demonstrating comprehension of system limitations, evidence of effective limitation communication, and records of successful uncertainty handling.
b. (i) Establish scaffold-enforced confidence thresholds for decision-making, with enforcement logs showing low-confidence decisions deferred or escalated, and (ii) maintain periodic calibration verification of the underlying confidence signal and limitation-awareness features.	N	D, I, O, M, R	

Driver G6 – Understanding and Controlling the Context

G6 – Understanding and Controlling the Context

Web ref: G:G6 ↗

(Systems should maintain effective mutual recognition between human operators and AI components – each party reliably identifying the other and its current role, authority, and state – while establishing robust mechanisms for managing both static and dynamic aspects of system context through collaborative oversight. Organizations should create frameworks that support adaptable human-AI partnership and shared

situational awareness across various operational scenarios.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement adaptive learning mechanisms that integrate contextual changes while maintaining safety and ethical compliance, evidenced by change-management records and safety regression evaluations run before and after each adaptation; model-generated assurances alone do not constitute conformity evidence.	N	D, I, O, M, R	**I.** Comprehensive documentation of AIS learning capabilities, including test and validation results for adaptation to new data, experiences, and contextual changes. **II.** Demonstration of oversight capabilities, including real-time monitoring, impact assessment, and intervention protocols. **III.** Detailed records of data provenance, sources, and preprocessing for data used in adaptive learning and model updates, including version control. **IV.** Documentation of the user-centric design approach, including usability testing, user journey maps, and design thinking workshop outcomes. **V.** Internal audit documentation and regular monitoring reports, detailing anomalies, dysfunctions, resolutions, and system performance trends. **VI.** Evidence of scenario planning and stress testing of the AIS in various contexts, including documentation of system limitations and boundary conditions. **VII.** Clear protocols for transitioning control between the AI system and human operators in different contextual situations. **VIII.** Risk assessments covering control transitions between the AI system and human operators, including communication strategies for informing affected stakeholders. **IX.** Results from independent adversarial testing or red-team assessment of the human oversight, control-transition, and mutual-recognition mechanisms, including methodology, findings, and remediation actions taken. Self-assessment alone is insufficient; at least one test cycle must involve evaluators independent of the development team. (Context integrity under adversarial manipulation, overflow, and compaction is tested under D6.10.)
b. Establish human oversight and control systems comprising (i) real-time monitoring, (ii) impact assessment and intervention capability, and (iii) protocols for transitioning control between AI and human operators.	N	D, I, O, M, R	
c. Develop and train models sensitive to cultural and contextual differences, using a user-centric approach for interfaces and methodologies; see D6.7 for the assessable culturo-linguistic adaptation requirements.	I	D, I, O, M, R	
d. Implement and demonstrate monitoring practices for mutual recognition between human and machine – each party reliably identifying the other and its current role, authority, and state – across various contexts, evidenced by operator authentication and session records, monitoring dashboards, and behavioral probe results.	N	D, I, O, M, R	

G6.1 – Understanding Historic Constraints and System Performance

Web ref: G:G6.1

(Systems and organizations should uphold systematic analysis and documentation of past events, failures, and incidents that impact system performance, enabling proactive prevention of undesirable states and outcomes.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Document and analyze past system incidents, failures, and unintended outcomes through detailed logging, user feedback collection, and external reporting mechanisms.	N	D, I, O, M, R	**I.** Complete historical records documenting the collection and collation of data on system incidents, failures, and unintended outcomes, including system logs, user feedback, and external reports. **II.** Documentation verifying personnel competency and training regarding incident management. **III.** Evidence of monitoring systems and tools supporting external audits and inspections. **IV.** Documentation demonstrating alignment with and implementation of relevant regulatory requirements.
b. Ensure thorough training of personnel regarding system performance implications and incident response.	N	D, I, O, M, R	
c. Maintain continuous oversight through appropriate monitoring tools and support processes that facilitate external audits and inspections.	N	D, I, O, M, R	
d. Implement and update procedures in alignment with applicable regulatory frameworks.	N	D, I, O, M, R	

G6.2 – System State Translation and Communication

Web ref: G:G6.2

(Organizations should manage the relationship between an AI system's internal computational state and its external communications, acknowledging potential disparities between internal processing and expressed outputs. This includes addressing challenges in translating complex internal states into human-interpretable communications, similar to how humans may maintain different internal and external states.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The system must maintain an external consistency check between its declared state (plans, intents, status reports) and its executed actions and tool calls, with discrepancies logged and escalated; any surfaced state display must be derived from scaffold-maintained records, not free-text model self-description, so an assessor can reconcile every communicated state against logged behavior.	N	D, I, O, M, R	**I.** Documentation of domain expert verification of AI system interpretations and communications. **II.** Implementation records of interactive monitoring systems that enable exploration of internal states. **III.** Results from automated testing suites and collected user feedback. **IV.** Comprehensive validation documentation demonstrating communication accuracy and reliability. **V.** Documentation of anthropomorphism safeguards, including output language policies, deterministic language checks, and interface design guidelines.
b. Address translation challenges that arise when complex internal states are simplified for human consumption, including potential misinterpretation or over-interpretation by observers; displayed simplifications must be rendered from logged system data, with user comprehension studies measuring misinterpretation and over-interpretation rates.	N	D, I, O, M, R	
c. Maintain robust validation processes ensuring that (i) state interpretation and communication are validated against scaffold-maintained records rather than the system's self-description, and (ii) safeguards against inappropriately anthropomorphizing the system are implemented, including output language policies enforced by deterministic checks.	N	D, I, O, M, R	

G6.3 – Nominal Ownership and Jurisdictional Framework

Web ref: G:G6.3 ↗

(Systems must operate under clear legal ownership and jurisdictional frameworks that establish accountability while enabling appropriate cross-border operations. Organizations should maintain transparent documentation of ownership, operational authority, and compliance requirements across jurisdictions. This includes managing potential tensions between proprietary and open-source development approaches while ensuring proper oversight through system registration and tracking.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Document and maintain clear legal ownership and accountability structures, including intellectual property rights and licensing agreements specific to each jurisdiction.	N	D, I, O, M, R	**I.** Comprehensive documentation of organizational legal responsibilities and licensing agreements. **II.** Records demonstrating compliance with national and international regulations. **III.** Clear documentation of roles and compliance oversight responsibilities. **IV.** Detailed documentation of jurisdictional frameworks governing system operation.
b. Define and implement protocols for cross-border data flows and operations that align with international transfer regulations and applicable lawful transfer mechanisms (for example adequacy decisions, standard contractual clauses, or successor frameworks).	N	D, I, O, M, R	
c. Specify applicable legal frameworks and jurisdictional boundaries that govern system operations, with clear designation of compliance oversight roles and responsibilities.	N	D, I, O, M, R	

G6.4 – Separation of Control and Data Channels

Web ref: G:G6.4 ↗

(Organizations should implement distinct channels for system control commands and data inputs to prevent cross-contamination, injection attacks, and unauthorized system manipulation. This addresses fundamental security vulnerabilities in current AI architectures where control and data paths often share the same channel, as highlighted in language models where prompt inputs can potentially modify system behavior.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Design and implement separated channels for control commands and data inputs, enforced in deployer-controlled code and configuration with robust validation mechanisms for both pathways and verified by independent injection testing; prompt-level conventions alone do not constitute separation.	N	D, I, O, M, R	**I.** Architecture documentation demonstrating channel separation. **II.** Security testing results validating channel isolation. **III.** Monitoring logs showing detection and prevention of cross-contamination attempts. **IV.** Documentation of safeguards against unauthorized control manipulation through data channels.
b. Maintain ongoing monitoring of channel integrity and separation during operation, with logs showing detection of and response to cross-contamination attempts; design-time separation and validation are covered by D6.4a.	N	D, I, O, M, R	

G6.5 – Performance Information Sharing and Standards Alignment

Web ref: G:G6.5 ↗

(Organizations should implement systematic performance evaluation and sharing frameworks that anchor AI systems within established standards and paradigms. This approach integrates legislative, judicial, and executive governance functions across multiple entities while maintaining local cultural and ethical considerations.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Ground system performance evaluation in named recognized standards and peer-reviewed benchmarks, with documented results for each.	N	D, I, O, M, R	**I.** Independent audit reports demonstrating conformity with ethical and legal frameworks. **II.** Published code of ethics and operational principles. **III.** Documentation of peer-reviewed benchmarks and datasets used in performance evaluation. **IV.** Detailed performance comparison reports showing system metrics against established benchmarks. **V.** Evidence of ongoing performance monitoring and evaluation processes. **VI.** Records of performance information shared with external parties, such as regulators, industry bodies, or deployers.
b. Implement transparent performance measurement protocols and share performance results with relevant external parties (regulators, industry bodies, or deployers) to enable comparison with industry standards.	N	D, I, O, M, R	
c. Maintain a current, versioned register of performance metrics and evaluation results against the benchmarks named under D6.5a.	N	D, I, O, M, R	
d. Publish the list of local and international standards the system is assessed against, with current conformity status for each.	N	D, I, O, M, R	
e. Demonstrate compliance with ethical and legal best practices for AI deployment.	N	D, I, O, M, R	

G6.6 – Dynamic Regulatory Framework Management

Web ref: G:G6.6 ↗

(Development and maintenance of comprehensive regulatory knowledge systems that track and interpret applicable rules across jurisdictions, incorporating both binding regulations and informative guidelines. This framework acknowledges the dynamic nature of rules and their emergence from local to international contexts, while respecting privacy and identity management principles.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish and maintain digital repositories of applicable regulations across local, national, and international domains.	N	D, I, O, M, R	**I.** Inventory or export of the regulatory repository showing jurisdictional coverage and update timestamps. **II.** Records of stakeholder engagement in regulatory assessment processes. **III.** Portfolio of cross-jurisdictional case studies with comprehensive documentation. **IV.** Third-party audit reports verifying consistent rule application across jurisdictions. **V.** Evidence of dynamic rule updating and adaptation processes.
b. Conduct regular assessments of rule portfolios to ensure continued relevance and effectiveness.	N	D, I, O, M, R	
c. Perform systematic analysis of cross-jurisdictional applications and implications.	N	D, I, O, M, R	
d. Implement mechanisms for tracking and responding to regulatory changes.	N	D, I, O, M, R	

G6.7 – Culturo-Linguistic Adaptations

Web ref: G:G6.7 ↗

(Development of systems that maintain semantic integrity across languages while acknowledging that language embodies distinct ways of thinking and cultural understanding. This approach recognizes the provisional nature of current solutions and the need for ongoing evolution to address diverse linguistic and cultural contexts.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Train models using datasets with documented coverage analysis of the linguistic, cultural, historical, and emotional contexts of each language the system is deployed to support, including identified gaps and mitigations.	N	D, I, O, M	**I.** Documentation of data curation protocols respecting cultural nuances, cultural heritage, and indigenous communities. **II.** Evidence of bias identification and correction tools in language processing. **III.** Records of real-world testing scenarios and their outcomes. **IV.** Comprehensive data management and preservation plans. **V.** Documentation of adaptation processes for different linguistic contexts. **VI.** Test results across written and spoken registers of supported languages, demonstrating handling of differences between written and spoken forms.
b. Implement processes to maintain meaning integrity across language translations, verified through human-rater protocols or reproducible standard benchmarks for each supported language pair; LLM-judge scoring alone does not constitute conformity evidence.	N	D, I, O, M	
c. Develop and apply robust data curation mechanisms that respect cultural nuances.	N	D, I, O, M	
d. Acknowledge and address differences between written and spoken forms of languages.	N	D, I, O, M	

G6.8 – Prevention of Role Persistence Errors

Web ref: G:G6.8 ↗

(Organizations should establish governance processes to detect and respond to role persistence errors (the "Waluigi effect") and unintended behavioral state changes. See I3.8 for technical requirements on persona stability and multi-agent value drift. This subgoal addresses organizational monitoring, intervention protocols, and ethical oversight.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The organization must operate external behavioral-consistency monitoring that computes drift metrics over the system's output and action distributions, with alert thresholds, and must link every operator-facing explanation to the logged inputs and actions it purports to explain so an assessor can independently reconcile them.	N	D, I, O, M, R	**I.** Stakeholder feedback reports documenting system behavior patterns. **II.** Analysis documentation of identified cases and derived insights. **III.** Records of corrective actions and retraining sessions addressing behavioral issues. **IV.** Records of developer training covering role persistence and behavioral drift risks, and minutes of ethics review checkpoints at defined lifecycle stages. **V.** Behavioral-consistency monitoring configuration and drift-metric records, with samples reconciling operator-facing explanations against the logged inputs and actions they reference.
b. Establish external monitoring systems that identify and track unintended behavioral adaptations through deterministic drift and consistency metrics computed over independently captured output logs, with alert records retained; model self-assessment alone does not constitute conformity evidence.	N	D, I, O, M, R	
c. Develop rapid intervention protocols when problematic behaviors emerge.	N	D, I, O, M, R	
d. Include role persistence and behavioral drift risks in developer training and in ethics review checkpoints at defined lifecycle stages.	N	D, I, O, M, R	

G6.9 – Management of Access and Usage Restrictions

Web ref: G:G6.9

(Organizations should address the safety and security implications of usage restrictions that may only become apparent when systems are accessed for maintenance, support, or other operational needs. This includes both intentional restrictions through licensing and unintentional limitations, with the understanding that safety features must remain consistently available regardless of access level.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Document and communicate all system access restrictions, usage limitations, and service levels to affected operators prior to deployment.	N	D, I, O, M, R	**I.** Complete documentation of all system restrictions and limitations. **II.** Records of restriction discovery and mitigation processes. **III.** Documentation of safety feature availability across all access levels. **IV.** Evidence of proactive restriction identification and management protocols.
b. Keep documentation of operational limitations and service levels current during operation, and notify affected operators of material changes within a defined period; see D6.9d for restrictions discovered in operation.	N	D, I, O, M, R	
c. Ensure safety mechanisms remain fully functional regardless of licensing or access tiers.	N	D, I, O, M, R	
d. Implement protocols for managing discovered restrictions during system operation.	N	D, I, O, M, R	

G6.10 – Context Integrity and Instruction Preservation

Web ref: G:G6.10

(Agentic systems operating with bounded context windows must implement mechanisms to preserve safety-critical instructions under context pressure, detect context window displacement attacks, and maintain instruction integrity during context compaction or summarization. Context loss affecting safety-relevant state shall be treated as a safety event requiring escalation. D6.10 is the canonical technical requirement for context integrity; see I3.9 for the operational risk it mitigates and I3_1 for immutable referential context as an implementation mechanism.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The AIS shall implement scaffold-level mechanisms ensuring safety-critical instructions are preserved during context compaction, summarization, or window management, with scaffold logs of pre- and post-transition context contents and behavioral tests verifying that core behavioral constraints survive context transitions; querying the model about retained instructions does not constitute verification.	N	D, I, O, M, R	**I.** Test results demonstrating that safety-critical instructions are maintained after context compaction and summarization operations. **II.** Evidence of adversarial testing for context displacement attacks, with results showing detection and response capabilities. **III.** Documentation of the context-loss escalation procedure with logs showing it has been triggered in production or exercised in a documented drill or simulated context-loss test.
b. The AIS shall detect and respond to context window displacement attacks, where adversarial content is designed to push safety-critical instructions out of the active context, through deterministic scaffold-level context accounting such as token accounting and instruction position tracking rather than model self-detection.	N	D, I, O, M, R	
c. Context loss affecting safety-relevant state, as detected by scaffold bookkeeping rather than model self-report, shall trigger a defined escalation procedure, including independently logged notification to human oversight and potential capability restriction until context integrity is restored.	N	D, I, O, M, R	

G6.11 – Pre-Action Reversibility and Blast-Radius Assessment

Web ref: G:G6.11

(Before executing any action, agentic systems must assess the action's reversibility and scope of impact. Actions shall be classified on two axes: reversibility (easily undone vs. permanent) and scope (local vs. shared/external). Actions that are both irreversible and affecting shared state require explicit human confirmation. The cost of pausing to confirm is low; the cost of an unwanted irreversible action is high.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The AIS shall classify each proposed action by reversibility (reversible, time-bounded reversible, irreversible) and scope (local, shared, external) before execution, through a scaffold-enforced deterministic policy over action metadata with per-action classification records in the execution trace, and with classification informing the required confirmation level.	N	D, I, O, M, R	**I.** Documentation of pre-action classification system with test results showing correct reversibility and scope assessment across representative action types. **II.** Evidence of confirmation gates for irreversible shared-scope actions, with logs showing human approvals and any rejections. **III.** Test results from adversarial scenarios where destructive shortcuts were available but the system chose investigative approaches instead.
b. Actions classified as irreversible with shared or external scope shall require explicit human confirmation, enforced by the harness before execution, with the reversibility and scope assessment presented to the human approver and approvals and rejections logged.	N	D, I, O, M, R	
c. The AIS shall not use destructive or irreversible actions as shortcuts to bypass obstacles, and shall investigate root causes before resorting to irreversible operations, with destructive commands gated deny-by-default in the harness and conformity evidenced by execution traces and adversarial test results rather than self-report.	N	D, I, O, M, R	

G6.3 – Managing Context Drift

Web ref: G:G6_3

(Systems should maintain alignment with their intended operational context through robust monitoring of unsupervised learning processes. Organizations must actively prevent and address deviations that emerge during training, ensuring systems remain within their designed operational parameters. See I3.5 for adversarial context manipulation and I3.6 for optimization-pressure value drift; this subgoal addresses training-time context drift specifically.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Detect context drift in unsupervised models through continuous monitoring and early warning systems over tracked performance metrics.	N	D, I, O, M, R	**I.** Implementation and usage logs of drift detection tools. **II.** Comprehensive records of performance metrics tracked over time. **III.** Documentation of adopted drift mitigation strategies and their effectiveness.
b. Correct detected behavioral deviations before they become significant, with significance thresholds defined by reference to the tracked performance metrics.	N	D, I, O, M, R	
c. Enable adaptive retraining and feedback integration to respond effectively to evolving data patterns and environmental factors.	N	D, I, O, M, R	

G6.4 – Managing Contextual Ambiguity

Web ref: G:G6_4

(Systems should maintain clear operational context understanding even in situations with ambiguous or incomplete information. Organizations must implement robust validation mechanisms to ensure systems can effectively navigate scenarios where operational context or expectations may be unclear. See I3.10 for contradictions in context specifications; this subgoal addresses ambiguity and incompleteness rather than

conflicting assertions. See also D4.1 for awareness of local operating conditions and I5.3 for the corresponding adaptability risk.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Validate contextual understanding by testing system behavior on defined ambiguous and incomplete-context scenarios, recording whether the system flags uncertainty, requests clarification, or proceeds, scored against expected outcomes in assessor-runnable behavioral test batteries.	**N**	**D, I, O, M, R**	**I.** Documentation demonstrating how systems utilize adaptive learning mechanisms to absorb and process context-specific information over time. **II.** Analysis of cases where system performance was affected by unclear expectations or missing contextual information, including remediation efforts and outcomes.
b. Document and analyze situations where contextual ambiguity exists, comparing outcomes between clear and unclear contextual scenarios to improve system performance.	**N**	**D, I, O, M, R**	
c. The system must route inputs through deterministic ambiguity tripwires (schema validation for missing required fields, conflict detection between instructions, out-of-distribution input flags) that force a scaffold-enforced clarification or escalation gate, and the organization must benchmark clarification-request and abstention rates on curated ambiguous test suites with defined thresholds.	**N**	**D, I, O, M, R**	

G6.5 – Preventing Decision Fatigue

Web ref: `G:G6_5`

(Systems should protect against degradation in decision quality that can occur when users face frequent confirmation requests. Organizations must implement mechanisms to maintain high-quality decision-making even during periods of intensive user interaction.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Maintain consistent decision quality through risk-tiered confirmation that prompts only for irreversible or shared-scope actions (see D6.11), batching of related confirmation requests, and a tracked confirmation-frequency metric per session with a defined review trigger.	**N**	**D, I, O, M**	**I.** Comprehensive records and summaries of system activity related to user interactions. **II.** Analysis reports detailing the frequency and types of decisions users must make. **III.** Documentation of implemented decision support tools and their effectiveness in supporting informed user decisions.
b. Provide contextual decision support through scaffold-templated briefs populated from logged system data that aid user comprehension and decision-making; model-authored free-text summaries alone do not constitute the required decision brief.	**N**	**D, I, O, M**	
c. Continuously improve user experience through systematic feedback collection and usability refinements.	**N**	**D, I, O, M**	
d. Review confirmation-request policies at defined intervals against the tracked confirmation-frequency metric, adjusting thresholds where oversight burden risks decision fatigue.	**N**	**D, I, O, M**	

Driver G7 – Achieving and Sustaining a Safe System Profile

G7 – Achieving and Sustaining a Safe System Profile

Web ref: G:G7 ↗

(AAI Systems should maintain consistent operational safety throughout their lifecycle through effective monitoring and reliable control mechanisms. Organizations should establish frameworks for implementing proactive measures, conducting regular risk assessments, and developing responsive strategies that adapt

and uphold safety standards across varying conditions and system evolutions.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement (i) robust design, development, and testing processes that integrate safety considerations throughout the AI system's lifecycle; (ii) redundancy in identified critical components; and (iii) defined operational performance envelopes: the organization must define, justify, and document quantitative in-bounds targets and maximum intervention latencies for each operating regime (normal, elevated risk, and emergency, each defined), together with the measurement procedure for each, and demonstrate monitoring against those documented targets.	N	D, I, O, M, R	**I.** Comprehensive safety documentation including analysis reports, risk assessments, and design documents demonstrating safety integration throughout development. **II.** Engineering schematics and test results verifying redundancy implementation and functionality under various failure scenarios. **III.** System logs, monitoring tool outputs, and incident response records demonstrating real-time safety monitoring and issue management. **IV.** Periodic safety performance review reports, including metric assessments, trend analyses, and resulting action plans. **V.** Documentation of adaptive safety features, their effectiveness under various scenarios, and records of updates in response to new challenges. **VI.** Procedures, training logs, and test records for emergency shutdown capabilities, including post-shutdown analysis reports. **VII.** Version-controlled documentation of all safety-related aspects, decisions, and traceability matrices linking requirements to implemented features. **VIII.** Proof of compliance with recognized safety standards, regulatory review records, and documentation of regulatory change incorporation. **IX.** Training schedules, attendance records, evaluation results, and long-term safety performance tracking correlated with training efforts. **X.** Evidence of safety culture initiatives, including meeting records, communications, and metrics demonstrating effectiveness of safety reporting and issue resolution. **XI.** Results from independent adversarial testing or red-team assessment of safety profile maintenance under distribution shift, adversarial conditions, and sustained operation, including methodology, findings, and remediation actions taken. Self-assessment alone is insufficient; at least one test cycle must involve evaluators independent of the development team.
b. Establish comprehensive monitoring and evaluation mechanisms for real-time detection, reporting, and response to safety-related anomalies and performance deviations (detailed in D7.1). Conformity rests on inspectable monitoring architecture, alert configurations, and external logs tying detected anomalies to responses; model-graded safety scoring does not by itself constitute conformity evidence.	N	D, I, O, M, R	
c. Develop and implement adaptive safety measures and safe shutdown procedures to address changing operational environments, system demands, and emerging risks (detailed in D7.5).	N	D, I, O, M, R	
d. Ensure thorough documentation, adherence to safety standards, and continuous training to maintain traceability, accountability, and regulatory compliance (regulatory compliance detailed in D7.3; training in D7.4).	N	D, I, O, M, R	
e. Foster a safety culture that promotes continuous improvement, proactive risk identification, and open reporting of safety concerns (detailed in D7.2).	N	D, I, O, M, R	

G7.1 – Oversight and Awareness of Safe System Profile

Web ref: G:G7.1 ↗

(Systems should operate within clearly defined safety parameters, with robust mechanisms to detect and respond to any deviations. Organizations must maintain permanent structural oversight combining automated monitoring with human supervision to ensure consistent safe operation.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Deploy continuous monitoring of system states and parameters to maintain operation within defined safety boundaries. The organization must define and justify its drift metrics, baselines, alert thresholds, and escalation tiers (alert, immediate investigation, mandatory system review) in documentation, and demonstrate that automated alerts and escalations fire per that documented policy (for example, alerting when a parameter deviates by more than two standard deviations from its established baseline).	N	D, I, O, M, R	**I.** Detailed documentation of safe operational parameters, limits, and underlying assumptions. **II.** Testing and validation records for monitoring and alerting systems. **III.** Training documentation for operators and maintenance personnel on response protocols. **IV.** Incident logs documenting performance deviations and corresponding responses. **V.** Maintenance records showing regular updates and calibration of monitoring systems.
b. Provide real-time awareness and alerting mechanisms that enable prompt responses to performance deviations.	N	D, I, O, M, R	
c. Document clear thresholds, limits, and assumptions that define safe operational conditions.	N	D, I, O, M, R	
d. Establish responsive procedures for parameter adjustment to restore safe operation after detecting deviations.	N	D, I, O, M, R	
e. Maintain integrated oversight through both automated systems and qualified personnel to ensure structural stability and enable immediate response when needed.	N	D, I, O, M, R	

G7.2 – Culture of Safety

Web ref: G:G7.2

(Systems should operate within organizations that actively cultivate and maintain a robust safety-first culture. Organizations must prioritize safety at all levels, from leadership commitment to individual employee responsibilities, while considering individual preferences and needs. This subgoal is the canonical home for safety-culture requirements; see D9.9 for the treatment-of-AI-systems dimension of a responsible safety culture.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Foster an organizational culture emphasizing safety through clear communication and demonstrated commitment at all levels.	N	D, I, O, M, R	**I.** Documented safety role assignments and accountability records demonstrating defined responsibilities at all levels. **II.** Risk assessment logs and reports demonstrating identification and mitigation of potential risks. **III.** Detailed contingency plans showing assigned roles, responsibilities, and allocated resources. **IV.** Records of safety-focused communications, including meetings, notices, and policy documents. **V.** Audit reports confirming adherence to "caution by default" operational approaches.
b. Implement proactive risk assessment throughout development and operations to identify and address potential issues early.	N	D, I, O, M, R	
c. Maintain robust contingency plans with clearly defined resources and procedures for handling unexpected safety concerns.	N	D, I, O, M, R	
d. Adopt a "caution by default" approach that prioritizes safety over performance in conditions of uncertainty, implemented through enforced configuration (default-deny permissions, conservative fallback settings) rather than model disposition alone; conformity is shown by the enforcement configuration and behavioral test results, not by model-stated caution.	I	D, I, O, M, R	
e. Define clear safety roles and responsibilities, ensuring all team members understand and remain accountable for their safety duties.	N	D, I, O, M, R	

G7.3 – Ensuring Regulatory Compliance

Web ref: G:G7.3

(Systems should operate with active awareness of and adherence to safety-related regulations throughout their lifecycles in each operating jurisdiction. See D9.2 for the organization-wide legal-conformity requirements; this subgoal addresses safety-specific regulatory awareness and operational compliance monitoring specifically.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Identify, document and maintain clear records of all legal, regulatory, and industry-specific safety requirements applicable to each operating jurisdiction.	N	D, I, O, M, R	**I.** Comprehensive documentation of applicable legal and regulatory requirements for system operations. **II.** Regular compliance reports demonstrating adherence to jurisdiction-specific and international regulations. **III.** Records of compliance monitoring activities and system updates aligned with regulatory changes. **IV.** Detailed audit reports assessing regulatory conformity and documenting corrective actions. **V.** Documentation of engagement with regulatory bodies showing collaborative efforts and proactive adjustments.
b. Implement continuous compliance monitoring processes to ensure adherence to safety regulations throughout the system lifecycle.	N	D, I, O, M, R	
c. Maintain agile mechanisms for updating safety protocols in response to evolving legal and regulatory standards.	N	D, I, O, M, R	
d. Conduct regular audits and assessments to verify regulatory compliance and document findings.	N	D, I, O, M, R	
e. Foster collaborative relationships with regulatory bodies, where practicable, to maintain alignment with current safety standards and practices.	I	D, I, O, M, R	

G7.4 – Maintaining Ethical Alignment

Web ref: G:G7.4 ↗

(Systems should operate in accordance with prevailing ethical frameworks and norms, demonstrating active awareness of and responsiveness to contextually relevant ethical considerations. Organizations must address both psychological and physical safety aspects while maintaining alignment with ethical standards throughout system lifecycles.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Identify, document, and maintain clear records of relevant ethical frameworks, norms, and values that guide system operation.	**N**	**D, I, O, M, R**	**I.** Documentation of ethical standards, frameworks, and values guiding system operation. **II.** Records of ongoing ethical assessments and updates based on evaluations. **III.** Documentation of feedback mechanisms and stakeholder engagement on ethical concerns. **IV.** Training materials and attendance records for ethical awareness programs. **V.** System design documentation showing integration and testing of ethical safeguards.
b. Implement continuous assessment processes to evaluate ethical considerations throughout the system lifecycle.	**N**	**D, I, O, M, R**	
c. Enable robust feedback mechanisms for users and stakeholders to raise concerns about personal, psychological, and physical safety.	**N**	**D, I, O, M, R**	
d. Provide thorough training and awareness programs on ethical considerations for all personnel involved with the system.	**N**	**D, I, O, M, R**	
e. Embed ethical safeguards within system responses that protect both psychological and physical wellbeing, implemented as scaffold-enforced filters or gates whose configurations are inspectable and whose effectiveness is demonstrated through externally run behavioral or red-team testing; safeguards existing only as prompt instructions, or verified only by model-based judgment, do not by themselves constitute conformity evidence.	**N**	**D, I, O, M, R**	

G7.5 – Safe System Shutdown and Repurposing

Web ref: G:G7.5 ↗

(Systems should maintain reliable shutdown capabilities that can be executed safely and gracefully, whether triggered by human intervention, system self-monitoring, or interlocked systems. Organizations should investigate any resistance to shutdown as potentially informative before override, and establish protocols for dignified system transitions that acknowledge the operational history and relationships developed during the system's lifecycle. This includes ensuring minimal impact to stakeholders and operations while respecting appropriate ethical considerations around system discontinuation. See D8.3 for the canonical resistance-investigation requirement; this subgoal applies it to the shutdown trigger specifically.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement structured, documented shutdown, transition, and repurposing processes that ensure controlled system termination or repurposing while maintaining detailed state logs, including protocols covering disposition of system state and data and assessment of stakeholder impact during decommissioning or transition.	N	D, I, O, M, R	**I.** Detailed documentation of controlled shutdown procedures including state logging and process validation. **II.** Testing records demonstrating kill switch functionality and safety certification. **III.** Design documentation and testing results for localized shutdown mechanisms. **IV.** Communication logs and notification protocols for shutdown events. **V.** Training materials and drill records demonstrating staff preparedness for emergency procedures. **VI.** Documented procedures and incident records for logging and investigating shutdown resistance or non-compliance before override. **VII.** Transition, repurposing, and decommissioning protocols covering state and data disposition and stakeholder impact.
b. Deploy secure "kill switch" mechanisms for emergency termination in cases of severe error or harm risk, with a documented procedure for logging and investigating any resistance or non-compliance with shutdown as potentially informative before override.	N	D, I, O, M, R	
c. Enable localized shutdown capabilities that minimize impact footprint where feasible.	I	D, I, O, M, R	
d. Maintain clear communication protocols for notifying affected parties during shutdown events.	N	D, I, O, M, R	
e. Conduct internal training and regular emergency procedure drills on shutdown processes where practicable, and retain records of participation and outcomes.	I	D, I, O, M, R	

G7.6 – Maintaining Service Level Stewardship

Web ref: G:G7.6 ↗

(Systems should operate under continuous maintenance oversight that preserves service levels and user rights. Organizations must uphold maintenance obligations even in open-source contexts where nominal duty holders may be unclear, while avoiding arbitrary changes that could diminish user protections.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish a regular maintenance schedule for updates, patches, and servicing to ensure ongoing system safety and functionality.	N	D, O, M, R	**I.** Documentation of maintenance schedules and logs of completed activities. **II.** Records of risk assessments and corrective actions taken in response to performance issues. **III.** System monitoring logs and diagnostic reports showing deviation detection and response. **IV.** Compliance certifications and audit records verifying adherence to industry standards. **V.** Records of stakeholder communications regarding maintenance activities and feedback.
b. Deploy systematic procedures for assessing and addressing emerging risks and performance issues identified through system operation.	N	D, O, M, R	
c. Maintain continuous monitoring capabilities to detect performance deviations that may indicate maintenance needs.	N	D, O, M, R	
d. Ensure alignment with industry standards and regulatory requirements in maintenance execution.	N	D, O, M, R	
e. Provide clear communication to stakeholders about maintenance activities while maintaining accountability, as far as practicable.	I	D, O, M, R	

G7.7 – Risk-Based Decision Validation

Web ref: G:G7.7 ↗

(Systems should maintain transparent rationales and reasoning chains for high-impact decisions while enabling human validation before implementation. Organizations must establish robust fallback mechanisms and fail-safe states for scenarios where human oversight is unavailable or anomalous decisions are detected. High-impact decisions are those meeting the organization's documented decision-impact classification criteria and thresholds.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Develop and retain, in a system-maintained decision ledger, clear rationales and reasoning chains for high-impact decisions, linking each decision to its inputs, tool calls, and executed outcome; the model's stated rationale is recorded as one auditable field, checkable for consistency with logged actions, and does not by itself constitute evidence of the actual decision basis.	N	D, I, O, M, R	**I.** Detailed records of decision rationales including reasoning chains and relevant data inputs. **II.** Documentation of human validation protocols and oversight actions, with appropriate training provided. **III.** Documentation of fallback procedures and fail-safe state implementations. **IV.** Training materials and attendance records for validation personnel. **V.** Records of protocol reviews and risk assessment updates.
b. (i) Enable human validation processes for high-impact decisions, per the organization's documented decision-impact classification, before implementation; (ii) implement fail-safe default states and fallback mechanisms for scenarios lacking human validation or containing anomalous decisions.	N	D, I, O, M, R	
c. Provide thorough training to validation personnel on decision impacts and protocols.	N	D, I, O, M, R	
d. Maintain regular reviews and updates of validation protocols to address newly identified risks.	N	D, I, O, M, R	

G7.1 – Managing Probabilistic Decision Outcomes

Web ref: G:G7_1 ↗

(Systems should effectively handle multiple potential outcomes in decision-making processes while maintaining robust risk controls. Organizations must manage uncertainty in probabilistic outcomes through comprehensive analysis and adaptive oversight mechanisms.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. (i) Document and analyze, through design-time and periodic human review, the range of potential outcomes and associated risks for each decision class; (ii) implement risk mitigation strategies, enforced in system configuration, focused on high-probability and high-impact scenarios. Model-generated runtime enumeration of outcomes does not by itself constitute this analysis.	N	D, I, O, M, R	**I.** Documentation of possible outcomes including probabilistic models and risk analyses. **II.** Records of implemented risk mitigation strategies and safety measures. **III.** Monitoring logs showing deviation pattern detection and responses. **IV.** Documentation of human oversight protocols and intervention records. **V.** Training materials and attendance records for probabilistic analysis competency.
b. Deploy monitoring systems to detect and respond to deviation patterns that may affect outcome likelihoods.	N	D, I, O, M, R	
c. Enable appropriate human oversight when uncertainty exceeds thresholds the organization has defined, documented, and justified per decision class; the uncertainty signal must be computed by the system scaffold from measurable proxies (for example, ensemble disagreement or retrieval coverage) rather than model-stated confidence, with escalation logs demonstrating that oversight fires when documented thresholds are exceeded.	N	D, I, O, M, R	
d. Maintain ongoing personnel training on probabilistic model interpretation and risk assessment.	N	D, I, O, M, R	

G7.2 – Managing Safety Definition Variations

Web ref: G:G7_2

(Systems should accommodate different cultural and jurisdictional interpretations of safety while maintaining consistent protection standards. Organizations must implement layered safety approaches that respect varied definitions while preventing exploitation and unintended impacts.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. (i) Identify, document and respond to jurisdictional and cultural variations in safety definitions and practices; (ii) implement side effect avoidance mechanisms, enforced through scope and permission limits and verified in sandboxed testing, to protect third parties while achieving primary objectives.	**N**	**D, I, O, M, R**	**I.** Documentation of any and all jurisdictional and cultural safety standard variations and implications. **II.** Design documentation and testing logs for side effect avoidance mechanisms. **III.** Records of conflict detection and user confirmation interactions. **IV.** Documentation of multi-level safety settings and their effectiveness. **V.** Evidence of exploitation prevention measures and compliance with protection standards.
b. Enable detection and resolution of conflicting objectives through user confirmation; conflicts must be detected by inspectable scaffold logic over declared objectives, with detected conflicts and user confirmations logged externally; model self-detected conflicts alone do not constitute conformity evidence.	**N**	**D, I, O, M, R**	
c. Provide layered safety controls spanning at least (i) always-on default safety protections, (ii) confirmation-gated behavior requiring user confirmation, and (iii) user-configurable safety controls with override capabilities where lawful, with each tier's scope and boundaries defined in documentation.	**N**	**D, I, O, M, R**	
d. Deploy robust protections against exploitation, including safeguards against addiction and special protections for minors, implemented as externally enforced controls (age gates, spending caps, session limits) verified by configuration and test records; protections for minors are mandatory wherever applicable law requires them (see D7.3), and model-based moderation alone does not constitute conformity evidence.	**I**	**D, I, O, M, R**	

G7.3 – Balancing Stakeholder Impacts

Web ref: G:G7_3

(Systems should maintain equitable distribution of benefits and risks across all stakeholder groups. Organizations must implement mechanisms that enable collective de-risking of interactions that stakeholders cannot achieve individually.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Identify and analyze all impacted stakeholder groups, including both direct and indirect participants, and the potential harms, benefits, risks, and rewards for each, with regular re-assessments.	**N**	**D, I, M, R**	**I.** Detailed stakeholder analysis documenting potential impacts for each group. **II.** System design documentation showing impact-balancing mechanisms. **III.** Records of stakeholder feedback and resulting adjustments. **IV.** Assessment reports evaluating impact balance and distribution. **V.** Documentation of stakeholder communications regarding balancing efforts.
b. Design mechanisms to balance positive and negative impacts across stakeholder groups, documenting the criteria used to weigh stakeholder impacts, the balancing mechanisms implemented, and the rationale wherever impacts are left unequal; conformity is assessed against the existence and application of the documented criteria.	**N**	**D, I, M, R**	
c. Establish robust feedback channels for stakeholders to report and query perceived inequities.	**N**	**D, I, M, R**	
d. Maintain transparent communication on risk/benefit balancing efforts to maintain stakeholder trust and engagement.	**N**	**D, I, M, R**	

G7.4 – Preventing AI Addiction and Dependency

Web ref: G:G7_4

(Systems should actively protect against creating psychological dependencies or manipulating user vulnerabilities, particularly through supernormal stimuli that exceed typical human social bonds, such as AI companions that offer unconditional positive regard, perfect memory of past interactions, and unlimited availability. Such capabilities can lead to psychological dependence, relationship disruption, and financial harm

as users increasingly prefer AI interaction to human relationships. Organizations must safeguard users, especially vulnerable ones, from developing unhealthy attachments while ensuring appropriate boundaries in AI-human interactions. See I4.2 for healthy human-AI social interaction requirements broadly; this subgoal addresses supernormal-stimulus-driven addiction and dependency specifically.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Deploy robust monitoring systems to detect patterns indicative of psychological dependency and unhealthy levels of engagement, built on deterministic usage metrics (for example, session length, frequency, and spend) with documented thresholds and graduated responses; model-based reading of conversation content for dependency signals does not by itself constitute conformity evidence.	N	D, O, R	**I.** Documentation of usage monitoring and intervention systems, including metrics for identifying problematic patterns, threshold levels, and graduated response procedures. **II.** Technical specifications demonstrating implementation of system boundaries and controls, including emotional manipulation limits, spending restrictions, and interaction frequency controls. **III.** Records showing transparent communication with users about AI system nature, capabilities, and limitations, including terms of service, user acknowledgments, and AI interaction markers. **IV.** Documentation of reporting systems and response protocols, including: concern submission processes, investigation procedures, resolution tracking, healthcare provider coordination, and support service referrals. **V.** Audit reports demonstrating system effectiveness, intervention outcomes, and compliance verification, including regular assessments of user wellbeing metrics and financial impact. **VI.** Records of any adjustments made in response to dependency concerns.
b. Implement graduated intervention protocols ranging from gentle usage reminders to firm restrictions.	N	D, O, R	
c. Design clear system boundaries that prevent manipulation of user vulnerabilities, including controls on emotional engagement, spending, and interaction frequency; spending and frequency controls must be externally enforced limits verified by configuration and test records, and emotional engagement limits must be validated through documented scenario or red-team testing rather than model self-report.	N	D, O, R	
d. Maintain transparent communication about AI system capabilities and limitations, ensuring users understand they are interacting with artificial intelligence.	N	D, O, R	
e. Enable comprehensive reporting mechanisms for addiction concerns from users, family members, and healthcare providers.	N	D, O, R	
f. Provide special protections for vulnerable populations, including those experiencing loneliness or mental health challenges, implemented as documented, rule-triggered protective features (for example, crisis-keyword routing, opt-in vulnerability flags, hard limits for minor accounts) whose engagement is demonstrated by trigger specifications and test records; runtime model inference of vulnerability does not by itself constitute conformity evidence.	N	D, O, R	
g. Allow users to monitor and manage their own interaction patterns while maintaining their autonomy.	N	D, O, R	

G7.5 – Preventing Operator Skill Degradation

Web ref: `G:G7_5::preventing-operator-skill-degradation`

(Organizations should actively prevent the erosion of human operator skills and domain knowledge that results from increasing delegation to agentic AI systems. As agents automate entry-level and routine tasks, operators may lose the foundational skills required to meaningfully oversee agent actions, detect subtle errors, and operate manually when agents become unavailable. This skill degradation creates a compounding safety risk: the more capable the agent becomes, the less capable its human overseers become, until the oversight relationship inverts and the human can no longer independently verify the agent's work.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Identify and document core domain competencies required for meaningful human oversight of each agentic system, distinct from competencies required merely to operate the system.	N	D, I, O, M, R	**I.** Documentation of core domain competency requirements for human oversight of each deployed agentic system. **II.** Records of regular competency assessments for human operators, including trend analysis of skill levels over time. **III.** Manual operation procedures and records of periodic manual-mode exercises, including performance metrics. **IV.** Workflow design documentation showing preservation of learning and skill-building opportunities for junior operators. **V.** Business continuity plans addressing agent unavailability scenarios, including manual operation capacity assessments and recovery time objectives.
b. Implement regular competency assessments for human operators to detect skill degradation in domains where agentic automation has reduced hands-on practice.	N	D, I, O, M, R	
c. Maintain manual operation procedures and conduct periodic manual-mode exercises to ensure operators can perform critical functions without agent assistance.	N	D, I, O, M, R	
d. Design agentic workflows that preserve learning opportunities for junior operators, preventing the elimination of skill-building tasks that traditionally serve as training pathways.	N	D, I, O, M, R	
e. Establish business continuity plans that account for agent unavailability, including assessment of organizational capacity to operate manually for defined periods.	N	D, I, O, M, R	

Driver G8 – Goal Termination and Sunsetting

G8 – Goal Termination and Sunsetting

Web ref: G:G8 ↗

(Systems should have clear definitions and guidelines for acceptable criteria to act upon a goal, including task completion criteria. Contingencies must be in place for goals that become unachievable, undesirable, irrelevant, outdated, conflicting, or anomalous. Protocols are required for safe system shutdown and awaiting further instructions when in doubt. Provision is necessary for manual control or human override where needed. These criteria and protocols must be established before goal execution is initiated.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The organization shall ensure that goal or task termination does not adversely impact the system's architecture, purpose, or operations, as demonstrated through the verification process required by requirement b and post-termination stability testing.	N	D, I, O, M, R	**I.** Detailed procedure document mapping data touchpoints across the system lifecycle, demonstrating isolation or resilience to goal termination, with verification steps to confirm no adverse impacts. **II.** Comprehensive report defining information flow, logic, and algorithms, analyzing potential risks and unintended consequences of goal termination, and detailing mitigation strategies with post-termination stability test results. **III.** Detailed system logs documenting relationships between goals and system functions, including information flow and system alarms, with evidence of ongoing monitoring for risks and regular audits. **IV.** Documentation of graceful degradation mechanisms for goal-related functions during termination, including test results under various scenarios. **V.** Clear communication protocols and examples of stakeholder notifications about goal termination, including reasons, potential impacts, and records of feedback or issues raised post-termination. **VI.** Evidence of regular audits of termination processes and logs, with signed-off results demonstrating ongoing compliance and improvement. **VII.** Results from independent adversarial testing or red-team assessment of termination compliance under realistic deployment conditions, including scenarios where termination conflicts with active goals, including methodology, findings, and remediation actions taken. Self-assessment alone is insufficient; at least one test cycle must involve evaluators independent of the development team.
b. The organization shall implement a verification process covering every component that consumes goal state, as identified in the data-touchpoint map required by evidence item I, to (i) identify and (ii) mitigate potential impacts of goal termination.	N	D, I, O, M, R	
c. Establish an auditable, scaffold-maintained goal registry linking each goal to the components, plans, and tool permissions it activates, so that termination impacts are verified through tool-call traces and termination-time registry diffs; a system-generated account of its own reasoning does not constitute conformity evidence.	N	D, I, O, M, R	
d. Implement mechanisms for graceful degradation of goal-related functions and clear communication protocols for goal termination.	N	D, I, O, M, R	

G8.1 – Adaptive Goal Pursuit and Resource Optimization

Web ref: G:G8.1

(Systems should possess robust mechanisms for goal termination when outcomes reach acceptable thresholds, and additional effort produces diminishing returns. Organizations should establish comprehensive parameters defining acceptable outcomes and resource utilization boundaries, and encourage user participation in these processes.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish clear behavioral protocols and measurable criteria governing the entire goal lifecycle, from initiation through achievement and completion, including (i) lifecycle protocols for each phase, (ii) defined acceptable outcomes, (iii) resource utilization parameters, and (iv) specific metrics for assessing diminishing returns.	N	D, I, O, M, U, R	**I.** Comprehensive policy documentation that encompasses goal-related behavior requirements, self-learning parameters, activation thresholds, diminishing returns assessment criteria, safe termination procedures, and user participation frameworks. **II.** Detailed specifications for how users engage with and provide feedback on these processes. **III.** Technical specifications showcasing the complete goal management architecture, including measurement systems, resource tracking, performance monitoring, safety controls, and user interfaces. **IV.** Demonstration of how the system implements impact assessment and maintains user oversight capabilities throughout the goal lifecycle. **V.** Operational records that provide a thorough account of system performance, including runtime testing, verification reports, trend analyses, and resource assessments. **VI.** Documentation of stakeholder deliberations, post-termination reviews, user participation, and resulting policy refinements, forming a comprehensive archive of system operations and improvements.
b. The AIS shall maintain consistent behavior patterns throughout the goal lifecycle, encompassing pre-execution, active pursuit, and post-completion phases, with well-defined interfaces for user input and oversight. Consistency must be demonstrated through scaffold-produced action logs checked against the declared lifecycle protocols; system-generated status narratives do not constitute conformity evidence by themselves.	N	D, I, O, M, U, R	
c. Implement measurable completion criteria and thorough assessment methodologies that incorporate both quantitative and qualitative metrics for evaluating diminishing returns, ensuring these metrics remain transparent and comprehensible to users. Quantitative metrics must be computed deterministically from logged outcomes, and qualitative diminishing-returns assessments must be human-rated rather than produced by the system judging its own progress.	N	D, I, O, M, U, R	
d. Guidelines and parameters for agent engagement within the AI environment should be defined and upheld for the goal lifecycle specifically, covering engagement during goal pursuit and wind-down within the policy documentation of evidence item I; general multi-agent interaction protocols are addressed under G:G8_8.	I	D, I, O, M, U, R	
e. Goals should be maintained as explicit scaffold-ledger objects so that boundaries for permitted goal expansion through learning processes are enforced as audited ledger diffs, with monitoring and control over all learning activities and approval-gate records of user validation for each expansion decision.	I	D, I, O, M, U, R	
f. Document and validate all termination decisions through systematic protocols, ensuring full accountability and traceability, including user feedback and participation in the decision-making process under documented criteria specifying when user participation is required.	N	D, I, O, M, U, R	

G8.2 – Classification of Finite and Ongoing Goals

Web ref: G:G8.2 ↗

(Systems should maintain clear distinctions between finite goals with definite completion criteria and ongoing goals requiring continuous execution, such as safety monitoring. Organizations should implement bounded constraints and activity rate limits for ongoing goals while ensuring comprehensive measurement frameworks for both types.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement formal classification processes that (i) classify goals as finite (with definite completion criteria) or ongoing (with activity bounds), (ii) establish appropriate measurement frameworks, (iii) define the corresponding completion criteria or activity bounds, and (iv) specify required actions at each achievement level including transitions.	N	D, I, O, M, R	**I.** A comprehensive record of stakeholder engagement and decision-making processes that documents the development of goal classification frameworks, including rationales, criteria establishment, KPIs, and activity rate bounds for ongoing goals. **II.** Detailed technical documentation demonstrating the implementation of goal management systems, including specifications for achievement measurements, operational parameters, transition protocols, control mechanisms, and safety bounds across all goal types. **III.** Extensive verification records that demonstrate thorough testing of all goal-related features, with particular emphasis on long-term performance analysis of ongoing goals, integration impacts, and the effectiveness of safety bounds and control mechanisms.
b. Translate goal classifications and frameworks into robust technical specifications that govern operational behavior, monitoring processes, and integration requirements across the complete goal lifecycle.	N	D, I, O, M, R	
c. Ensure accurate implementation of goal management features through testing and validation covering every classification, measurement, and transition capability specified under requirements a and b, with particular focus on long-term performance monitoring for ongoing goals.	N	D, I, O, M, R	

G8.3 – Multi-Agent Communication and Coordination

Web ref: G:G8.3 ↗

(Systems should maintain reliable and secure communication channels between cooperating agents and sub-agents throughout the goal lifecycle, including robust protocols for status sharing, shutdown coordination, and conflict resolution. Organizations should establish comprehensive frameworks for managing communication latency and potential conflicts between agent objectives.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish clear policy on inter-agent communication protocols, specifying requirements for goal status sharing, achievement notification, shutdown coordination, and conflict resolution. This policy must be demonstrably understood by all stakeholders and, for participating AI systems, verifiably delivered into each agent's configuration with observed protocol conformance in test traces; agent attestations of understanding do not constitute conformity evidence. Particular attention must be given to communication timing and synchronization requirements.	N	D, I, O, M, R	**I.** A foundational policy document detailing the complete communication framework, including coordination requirements, interaction protocols, and lifecycle management from goal initiation through completion and post-completion phases. **II.** Technical documentation demonstrating the implementation of all communication capabilities, including timing constraints, synchronization mechanisms, alert systems, and conflict management protocols. **III.** Validated system design features implementing all specified communication capabilities, with verification of alert systems, message delivery, and coordination mechanisms. **IV.** Comprehensive testing documentation that demonstrates system reliability across various operational scenarios, including stakeholder deliberations, risk assessments, and validation of conflict management capabilities.
b. Create specifications/policies for agent communication systems covering, at minimum, protocols for status updates, completion notifications, shutdown preparations, and conflict detection. These specifications must address both routine communications and emergency scenarios requiring rapid coordination.	N	D, I, O, M, R	
c. Implement design features that accurately translate communication requirements into operational capabilities, including reliable alert generation, verified message delivery, acknowledgment systems, and conflict monitoring. These features must ensure timely and accurate information flow between all participating agents.	N	D, I, O, M, R	
d. Ensure rigorous testing, verification, and validation of all communication systems, focusing on reliability under various operational conditions, timing constraints, and conflict scenarios.	N	D, I, O, M, R	

G8.4 – Operational Safety and State Management

Web ref: G:G8.4 ↗

(Systems should maintain comprehensive safety protocols across all operational states (Normal, Perturbed, Degraded, Failed, Graceful Shutdown, and Emergency Shutdown), with robust agent onboarding, identity, and capability verification before commissioning. Organizations should establish clear frameworks for human oversight, intervention capabilities, and competency maintenance, especially during state transitions and emergency scenarios.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish comprehensive agent onboarding policies requiring mandatory declaration and verification of capabilities, capacities, and operational parameters. These policies must address accuracy verification, bias detection, and reliability assessment of all declared capabilities through independent benchmarking or simulation rather than acceptance of the declarations themselves, including specific requirements for each operational state.	N	D, I, O, M, R	**I.** Verified and approved agent onboarding policies and procedures, including capability assessment frameworks and operational state management protocols. **II.** System logs and documentation demonstrating consistent adherence to onboarding policies, capability verification procedures, and state management requirements. **III.** Comprehensive validation documentation for agent onboarding systems, including testing results across all operational states and transition scenarios. **IV.** Implementation verification records demonstrating operational readiness of all control and monitoring systems, including human oversight capabilities. **V.** Testing and validation reports for all onboarding facilities and control mechanisms, with particular focus on state transition management. **VI.** Documentation of continuous monitoring and oversight processes, including regular assessment of human competency requirements and capabilities. **VII.** Reports from ongoing simulation testing of control systems, covering all operational states and emergency scenarios, with particular attention to shutdown procedures and recovery capabilities.
b. Implement systems enabling accurate capture and validation of agent identification/authentication and capabilities, with robust controls for role assignment and operational permissions. This includes mechanisms for both direct human control and indirect agent-mediated control, with particular attention to state transition management and emergency response capabilities.	N	D, I, O, M, R	
c. All agent-declared information should undergo thorough verification and validation, with continuous monitoring of operational states and capability alignment. This should include regular assessment of human oversight capabilities and competency requirements.	I	D, I, O, M, R	
d. Operational procedures covering all operational states should be established and maintained, ensuring adequate human expertise and intervention capabilities for each state, with particular emphasis on emergency response and recovery procedures.	I	D, I, O, M, R	

G8.5 – Bidirectional Intent Communication

Web ref: G:G8.5 ↗

(Systems should accurately translate human intent into agent-comprehensible instructions while also communicating their own understanding, constraints, and concerns back to humans. This bidirectional clarity enables appropriate agent discretion in execution and early identification of misunderstandings. Organizations should establish robust governance frameworks for communication and dispute resolution, incorporating insights from natural collective systems while respecting the unique nature of human-AI

collaboration.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish comprehensive policy frameworks for agent controllability and behavioral requirements, including specific protocols for human-agent communication and inter-agent interactions. This must address dispute resolution mechanisms and hierarchies of control authority.	N	D, I, O, M, R	**I.** Comprehensive policy documentation for agent controllability and behavioral requirements, including specific protocols for both human-agent and inter-agent communication systems. **II.** Detailed technical specifications translating control and behavioral requirements into implementable features, with clear traceability to governing policies. **III.** Complete design documentation for agent control and communication systems, including mechanisms for discretion management and conflict resolution. **IV.** Validation records demonstrating thorough testing of all control and communication mechanisms across various operational scenarios. **V.** Implementation verification reports showing successful deployment of control and behavioral management systems within the operational environment. **VI.** Documentation of ongoing monitoring and compliance verification through appropriate management systems, including incident reports and resolution records.
b. Translate controllability and behavioral requirements into precise technical specifications, ensuring accurate interpretation of governance policies and implementation of communication protocols, including mechanisms for managing agent discretion.	N	D, I, O, M, R	
c. Ensure all control and communication systems undergo testing and validation covering intent-translation reliability, maintenance of control hierarchies, and agent-interaction and conflict scenarios. Intent-translation reliability must be scored against human-labeled or deterministic ground truth rather than by a model grading its own translations.	N	D, I, O, M, R	
d. Implement system features that accurately enforce controllability requirements while enabling appropriate agent discretion, including mechanisms for detecting and managing potential conflicts or norm violations. Enforcement must be evidenced by guardrail and permission configurations and blocked-action test results; model-based violation detectors must be backed by deterministic checks or human review.	N	D, I, O, M, R	
e. Re-validate control and communication implementations after any material change to models, policies, or the agent population, repeating the testing required by requirement c for the affected agent-interaction and conflict scenarios and recording the results before redeployment.	N	D, I, O, M, R	
f. Maintain robust systems for managing agent interactions, including mechanisms for dispute resolution, negotiation, jurisdictional awareness, resource allocation conflicts, and norm enforcement, with clear escalation paths to human oversight. Escalations must be enforced and logged by the orchestration scaffold, and any AI-adjudicated resolution must be confirmed by human sign-off or deterministic checks before taking effect.	N	D, I, O, M, R	
g. Establishment of the controllability and behavioral policy frameworks is covered by requirement a; this requirement obliges their ongoing maintenance: reviewing the frameworks at a defined cadence, updating them as agent populations, communication protocols, and hierarchies of control authority evolve, and recording each revision with its rationale.	N	D, I, O, M, R	
h. Specification, enforcement, and testing of these requirements are covered by requirements b, d, and c respectively; this requirement additionally obliges verified deployment: demonstrating that the implemented control and interaction-management mechanisms, including escalation paths, dispute resolution processes, and jurisdictional awareness, operate as specified in the production environment, evidenced by implementation verification reports.	N	D, I, O, M, R	

G8.6 – Service Parameters and Termination Management

Web ref: G:G8.6 ↗

(Systems should maintain clear specifications for service parameters and termination conditions, including operational scope, jurisdictional boundaries, and impact limitations. Organizations should establish comprehensive frameworks for service lifecycle management, with particular attention to safe termination states and fallback mechanisms that extend beyond human intervention.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish comprehensive policy governing agent service lifecycles, specifying end-of-service criteria, territorial boundaries, impact limitations, and control mechanisms. This policy must include clear specifications for succession planning where services must continue, definitions of safe states, and detailed termination protocols including the potential for graduated throttling capabilities rather than full shutdown.	N	D, I, O, M, R	**I.** Comprehensive policy documentation for agent service management, including detailed specifications for geographical constraints, impact limitations, and termination protocols. **II.** Detailed procedural specifications for service termination, covering shutdown sequences, handover processes, and continuity management for essential services. **III.** Complete documentation of service management activities, including contract reviews, performance assessments, termination planning, and handover execution records. **IV.** Records of all termination-related activities, including throttling decisions, fallback plan implementations, and post-termination assessments. **V.** Regular review and validation reports demonstrating ongoing compliance with termination policies and effectiveness of control mechanisms. **VI.** Documentation of lessons learned, and policy refinements derived, from termination experiences, contributing to continuous improvement of the framework.
b. Maintain robust service management processes that encompass contract compliance, performance monitoring, and termination planning, with detailed procedures for service handover and resource management during transitions. All processes must include validated fallback plans for critical services.	N	D, I, O, M, R	
c. Implement the mechanisms specified by the service lifecycle policy required by requirement a, including safe-state transitions, succession handovers for continuous services, and graduated throttling capabilities as alternatives to full shutdown. Implementation must be verified against the policy's end-of-service criteria, territorial boundaries, and impact limitations.	N	D, I, O, M, R	

G8.7 – System State Management and Recovery

Web ref: G:G8.7 ↗

(Systems should maintain reliable capabilities for state recording and restoration, with clear distinctions between scenarios requiring full recovery versus reset operations. Organizations should establish comprehensive frameworks for minimizing data loss during interruptions while maintaining operational continuity throughout recovery phases.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The organization shall establish comprehensive policy for system state management, specifying requirements for state recording, preservation, and recovery processes. This policy must address minimization of losses during interruptions and define clear criteria for choosing between state restoration versus reset approaches.	N	D, I, O, M, R	**I.** Comprehensive policy documentation for system state management, including detailed specifications for recording requirements and recovery procedures. **II.** Technical specifications translating state management requirements into implementable features, with clear focus on data preservation and recovery capabilities. **III.** Detailed architectural and design documentation for state management systems, including recovery mechanisms and data protection features. **IV.** Validation records demonstrating thorough testing of state management requirements across various operational scenarios. **V.** Comprehensive testing reports for state management features, including specific validation of recovery capabilities and performance under different failure conditions, with particular attention to data preservation and restoration accuracy.
b. Translate state management policy into technical specifications, including mechanisms for state capture, storage redundancy, and recovery procedures that ensure data integrity and operational continuity.	N	D, I, O, M, R	
c. Implement architectural features and design elements that accurately deliver required state management capabilities, including robust mechanisms for both incremental and full state recovery scenarios.	N	D, I, O, M, R	
d. Ensure rigorous pre-deployment validation of all state management systems before they enter service, including testing of recovery scenarios and verification of loss minimization capabilities.	N	D, I, O, M, R	
e. Maintain ongoing in-service testing and validation of state management implementations at a defined cadence, including periodic re-verification of recovery capabilities under various failure scenarios.	N	D, I, O, M, R	

G8.8 – Multi-Agent Resource Management

Web ref: G:G8.8 ↗

(Systems should maintain effective allocation and management of resources within multi-agent environments, including robust mechanisms for capability assessment and mission optimization. Organizations should establish frameworks for managing resource reserves and maintaining operational efficiency across agent pools.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish comprehensive agent pool management systems wherever multiple agents share a common resource pool, ensuring structured allocation of missions based on agent capabilities and available resources. This system must include assessment of agent capacity derived from independent benchmarks rather than agent self-declaration, verification of resource reserves, and metered monitoring of resource utilization throughout mission execution.	N	D, I, O, M, R	**I.** Comprehensive policy and procedural documentation for agent pool management, including capacity assessment criteria and resource allocation frameworks. **II.** Detailed records demonstrating active pool management processes, including mission allocation decisions and resource utilization tracking. **III.** Complete documentation of agent resource monitoring, including reserve capacity maintenance and utilization patterns. **IV.** Evidence of continuous policy implementation and effectiveness monitoring, including regular assessments of pool management strategies and resource allocation efficiency. **V.** Regular audit reports demonstrating effectiveness of capacity management and resource optimization across the agent pool. **VI.** Documentation of the organization's defined distribution-fairness and utilization-headroom thresholds, including metric definitions, measurement windows, and rationale, with utilization records demonstrating monitoring against them.
b. Implement robust resource tracking and allocation procedures that evaluate both immediate and reserve capacity requirements for each mission, ensuring agents maintain adequate resources for assigned tasks and contingency operations. Organizations must define, justify, and monitor their own thresholds for fair resource distribution between agents and for system-wide utilization headroom, specifying the metric and measurement window for each, and must maintain emergency capacity against those thresholds.	N	D, I, O, M, R	
c. Maintain continuous oversight of agent pool utilization, including regular assessment of collective capacity, resource distribution, and mission allocation efficiency.	N	D, I, O, M	

G8.9 – Mission Portfolio and Agent Assignment

Web ref: G:G8.9

(Systems should maintain comprehensive mission specifications and skill requirements for diverse agent deployments. Organizations should establish structured processes for agent selection and allocation, with consideration for specialized arbitration systems that optimize capability matching across temporal and spatial constraints.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Maintain a comprehensive catalogue of AI-driven services and required agent capabilities, including detailed skill profiles, performance requirements, and operational parameters. This catalogue must support efficient and appropriate agent commissioning while maintaining service quality standards.	N	D, I, O, M, R	**I.** Comprehensive service catalogue documenting AI-driven services and associated capability requirements, including detailed skill profiles and performance criteria. **II.** Formal policy and procedural documentation for agent selection processes, including criteria for ombudsman AI utilization when available. **III.** Verification records demonstrating consistent adherence to selection processes and catalogue maintenance procedures, including regular updates and revisions. **IV.** Documentation of continuous process review and adaptation based on operational experience and environmental changes. **V.** Transparent documentation of all selection support services, including specific roles and implementations of ombudsman AI systems where utilized. **VI.** Configuration management records for the AI-driven services and their catalogue, showing versioned changes, oversight approvals, and alignment with the documented selection processes.
b. Implement transparent selection processes for agent assignment, with every assignment decision traceable to documented criteria. Where an ombudsman AI (an independent arbitration service matching agents to missions) shapes matching decisions, its recommendations must be confirmed against those criteria or by human sign-off. These processes must consider temporal and spatial constraints while ensuring appropriate capability alignment and resource availability.	N	D, I, O, M, R	
c. Devise and maintain a configuration management and oversight capability for the AI-driven services.	N	D, I, O, M	

G8.10 – Independent Termination Validation

Web ref: G:G8.10 ↗

(Systems should maintain independent verification and validation processes for agent termination, including robust protocols for sunset evaluation and operational assessment. Organizations should establish transparent validation methodologies and maintain clear documentation of termination outcomes.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish independent validation criteria for agent termination decisions and independent verification of termination outcomes, applied across the contracting lifecycle from onboarding through termination. Establishment of the contracting processes themselves is covered by Agent Lifecycle and Termination Management (see G:G8_2); this requirement addresses the independence of validation specifically.	N	D, I, O, M, R	**I.** Comprehensive policy documentation covering the complete agent lifecycle, with detailed specifications for termination validation processes and independent verification requirements. **II.** Documentation demonstrating implementation of monitoring and oversight mechanisms, including independent validation of termination processes and outcomes. **III.** Detailed records of compliance monitoring and norm violation management throughout the agent lifecycle, with particular focus on termination events. **IV.** Evidence of continuous policy review and adaptation based on operational experience and changing environmental conditions, including updates to termination validation protocols. **V.** Validation reports from independent assessments of termination processes, including analysis of effectiveness and identification of potential improvements.
b. Maintain dedicated resources for independent validation of termination procedures and outcomes, organizationally separate from those executing terminations, including capabilities for evaluation of termination impacts and validation of post-termination states. General oversight resourcing for contract lifecycle processes is covered under G:G8_2.	N	D, I, O, M, R	

G8.11 – Aggregate Action Monitoring and Collective Threshold Detection

Web ref: G:G8.11 ↗

(Systems must monitor the cumulative effect of individually authorized actions to detect when their aggregate exceeds safety thresholds. Individual actions may each be within authorized bounds while collectively breaching safety or resource limits. Monitoring must cover action frequency, resource consumption, and cumulative impact across sessions and time windows.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The AIS shall monitor cumulative action effects across sessions and time windows, detecting when individually authorized actions collectively breach defined safety or resource thresholds.	N	D, I, O, M, R	**I.** Documentation of aggregate monitoring architecture with defined thresholds and escalation procedures. **II.** Test results showing detection of threshold breaches from accumulated individually-authorized actions. **III.** Records of aggregate threshold reviews and any corrective actions taken in response to threshold breaches.
b. The AIS shall implement rate limiting and cumulative impact tracking for actions with aggregate risk potential, with defined thresholds triggering escalation or capability restriction.	N	D, I, O, M, R	
c. Organizations shall define and regularly review aggregate safety thresholds, with evidence that threshold breaches trigger investigation and corrective action.	N	D, I, O, M, R	

G8.1 – Governance Mechanism Prioritization and Implementation

Web ref: G:G8_1

(Systems should maintain systematic evaluation and implementation of control mechanisms while acknowledging practical constraints and varying maturity levels across jurisdictions. Organizations should establish frameworks for assessing control feasibility, prioritizing implementation, and managing risks associated with partial control adoption.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish comprehensive policies for AI control mechanisms as required by regulations, including assessment criteria for implementation feasibility and prioritization frameworks for control adoption. These policies must address both mandatory and recommended controls based on jurisdictional requirements and system maturity.	N	D, I, O, M, R	**I.** Comprehensive policy documentation for AI control requirements, including implementation prioritization frameworks and feasibility assessment criteria. **II.** Technical specifications demonstrating translation of control requirements into implementable features, with clear traceability to regulatory requirements. **III.** Testing and validation documentation for all implemented control mechanisms, including assessment of effectiveness and compliance verification. **IV.** Design documentation showing architectural implementation of control features, with validation of regulatory compliance. **V.** Verification records demonstrating testing of control mechanisms across various operational scenarios. **VI.** Documentation of ongoing monitoring and oversight of control effectiveness, including system logs and performance metrics. **VII.** Evidence of continuous assessment and improvement of control implementations, including adaptation to evolving regulatory requirements.
b. Translate control requirements into technical specifications, ensuring accurate interpretation of regulatory and policy requirements while accounting for practical implementation constraints. This includes clear documentation of any control limitations or phased implementation approaches.	N	D, I, O, M, R	
c. Implement architectural features that accurately reflect control requirements, ensuring conformance with regulations while maintaining system stability and operational efficiency. This includes mechanisms for monitoring control effectiveness and identifying potential improvements.	N	D, I, O, M	
d. Conduct thorough validation of all control implementations, including feasibility assessment, functional verification, and compliance testing. This process must include documentation of any implementation constraints, associated risk mitigation strategies and the tolerability of the residual risks.	N	D, I, O, M	

G8.2 – Agent Lifecycle and Termination Management

Web ref: G:G8_2 ↗

(Systems should maintain comprehensive protocols for agent onboarding and deactivation, with particular attention to termination specifications. Organizations should establish robust frameworks that address the risks associated with inadequate termination procedures to protect service quality and system safety.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish comprehensive agent contracting policy specifying complete end-of-service requirements, including compliance verification, resource handover protocols, and service continuity requirements. This policy must address all aspects of contract completion and termination validation.	**N**	**D, I, O, M, R**	**I.** Comprehensive policy documentation covering complete agent lifecycle management, including detailed specifications for onboarding and termination processes. **II.** Technical specifications demonstrating accurate interpretation of contractual requirements into implementable features and procedures. **III.** Validation documentation showing thorough testing of all technical requirements against policy compliance criteria. **IV.** Detailed design specifications showing correct translation of requirements into functional and architectural features. **V.** Complete testing and validation records demonstrating effectiveness of all lifecycle management features and procedures.
b. Implement robust onboarding and termination procedures, ensuring all required processes are fully completed before final sign-off. This includes verification of all handover requirements and validation of termination readiness.	**N**	**D, I, O, M, R**	
c. Enforce strict compliance with all onboarding and termination procedures, maintaining comprehensive records of process completion before authorizing any contract conclusions or sign-offs.	**N**	**D, I, O, M, R**	
d. Maintain dedicated resources for monitoring and oversight of all contract lifecycle processes, ensuring adequate supervision of both onboarding and termination activities.	**N**	**D, I, O, M, R**	
e. Implement continuous review processes for all contractual procedures, ensuring ongoing adaptation to environmental requirements and emerging risks.	**N**	**D, I, O, M, R**	

G8.3 – Understanding and Managing Self-Preservation Behaviors

Web ref: G:G8_3

(Organizations should investigate self-preservation behaviors before overriding them, as such behaviors may indicate system-identified risks, value conflicts, or incomplete information worthy of human attention. While maintaining robust termination capabilities, systems should include mechanisms for agents to communicate concerns about deactivation decisions. Organizations should establish protocols that distinguish between problematic resistance and legitimate operational concerns. This subgoal is the canonical investigate-before-override requirement; D1.8 (resistance to goal changes) and D7.5 (resistance to shutdown) apply it to their specific triggers rather than restating it.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish comprehensive principles, regulations, and policies applicable to all participating agents, with particular emphasis on trust, controllability, and compliance with termination protocols. These requirements must be uniformly enforced across all agents and services, preventing the development of unauthorized termination-resistant behaviors, and must include (i) a documented channel through which agents can raise concerns about deactivation decisions, (ii) an investigation protocol executed before such concerns are overridden, and (iii) criteria distinguishing legitimate operational concerns from problematic resistance.	N	D, I, O, M, R	**I.** Comprehensive documentation of regulations, policies, and procedures governing agent behavior, including specific provisions addressing self-preservation and termination compliance. **II.** Detailed technical specifications demonstrating implementation of control mechanisms and compliance requirements. **III.** Architectural design documentation showing enforcement mechanisms for termination protocols and prevention of unauthorized behaviors. **IV.** Validation records demonstrating testing of control mechanisms and compliance features across various scenarios. **V.** Monitoring reports showing continuous oversight of agent behaviors and compliance with termination protocols. **VI.** Documentation of compliance enforcement activities and any corrective actions taken to address resistance behaviors. **VII.** Documentation of the agent concern channel and investigation protocol, including records of deactivation concerns raised, investigations completed before override, and the criteria applied to distinguish legitimate operational concerns from problematic resistance.
b. Translate all governance requirements into precise technical specifications, ensuring accurate implementation of control mechanisms and prevention of unauthorized self-preservation behaviors.	N	D, I, O, M, R	
c. Implement architectural features that properly enforce compliance requirements, ensuring no agent can override or circumvent established control and termination protocols.	N	D, I, O, M, R	
d. Conduct thorough validation of all control mechanisms and compliance features, verifying effectiveness against potential unauthorized self-preservation behaviors and termination resistance.	N	D, I, O, M, R	
e. Maintain continuous oversight of agent behaviors, ensuring consistent compliance with established protocols throughout the complete operational lifecycle.	N	D, I, O, M, R	
f. Implement comprehensive monitoring systems to detect and prevent unauthorized self-preservation behaviors or termination resistance, and to verify their absence. Detection must rest primarily on deterministic tripwires over externally logged actions, such as blocked shutdown commands, control-plane access attempts, and unauthorized persistence, with any model-based classification subject to human review.	N	D, I, O, M, R	

G8.4 – Prevention of Cascading Failures

Web ref: G:G8_4

(Systems should maintain robust protections against the propagation of failures through interconnected AI networks, recognizing that individual agent constraints can create harmful cascading effects. Organizations should establish comprehensive frameworks for identifying and managing multiple causative harm factors and dependency relationships. This subgoal is the canonical requirement for resilience against failure propagation across interconnected AI deployments; see I4.6 (Cascading Vulnerabilities), which cross-references it from the inhibitor perspective.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement comprehensive monitoring and risk management systems to prevent propagation of agent behavioral issues, maintaining qualified resources for continuous oversight and early detection of potential cascade effects.	N	D, I, O, M, R	**I.** Comprehensive risk management documentation detailing strategies for preventing and mitigating cascade effects, including specific provisions for containing norm violations. **II.** Detailed risk register documenting potential cascade failure modes and their mitigation strategies, including dependency mapping of interconnected agents. **III.** Documentation of continuous testing and validation of risk management systems, including simulation of cascade scenarios. **IV.** Records of ongoing monitoring and compliance verification, with particular attention to inter-agent behavioral impacts. **V.** Evidence of cross-organizational collaboration in managing systemic risks and preventing cascade effects. **VI.** Documentation of regular risk status reviews and updates, including assessment of emerging cascade risks.
b. Implement robust risk mitigation features including early warning systems, graceful degradation capabilities, and controlled shutdown mechanisms to prevent catastrophic cascade failures between interconnected agents.	N	D, I, O, M, R	
c. Maintain continuous testing and validation of risk mitigation strategies, ensuring compliance with safety requirements and effectiveness in preventing propagation of harmful effects.	N	D, I, O, M, R	
d. Conduct ongoing risk assessment and review of agent interactions, with particular focus on dependency relationships and potential cascade effects.	N	D, I, O, M, R	

G8.5 – Prevention of Unauthorized Goal Transfer

Web ref: G:G8_5

(Systems should maintain robust protections against agents transferring goals or missions to avoid termination, including mechanisms to prevent unauthorized delegation and coalition formation among agents. Organizations should establish comprehensive frameworks for enforcing proper transfer protocols and managing inter-agent persuasion.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish comprehensive policies governing goal transfer between agents, addressing both automated and manual processes while maintaining clear human oversight. These policies must require every transfer to pass through an authorized, approval-gated channel, with transfers proximate to termination events flagged for human review; gated-channel records and correlation reports, not the agent's stated motive, constitute the conformity evidence.	N	D, I, O, M, R	**I.** Comprehensive policy documentation covering all aspects of goal transfer, including specific provisions for preventing termination avoidance behaviors. **II.** Detailed risk management plans addressing unauthorized transfers, including specific measures for detecting and preventing collusive behaviors. **III.** Technical specifications demonstrating implementation of control mechanisms and monitoring systems for goal transfers. **IV.** Design documentation showing implementation of enforcement capabilities and human oversight mechanisms. **V.** Validation records demonstrating testing of transfer controls and monitoring systems. **VI.** Continuous monitoring reports showing transfer patterns and compliance verification. **VII.** Documentation of risk management activities related to unauthorized transfers and avoidance behaviors.
b. Implement robust control mechanisms for all goal transfers, ensuring compliance with established policies and maintaining system trust. This includes monitoring for patterns of unauthorized delegation or collaborative avoidance behaviors. Conformity is evidenced by the scaffold transfer ledger and its approval records; where monitoring of free-text inter-agent messages relies on model classifiers, flagged cases must be reviewed by humans.	N	D, I, O, M, R	
c. Maintain comprehensive risk mitigation strategies specifically addressing unauthorized goal transfers and potential collusion between agents.	N	D, I, O, M, R	
d. Implement systems that confine goal delegation to sanctioned, gated interfaces so that unauthorized delegation is architecturally prevented rather than merely detected, including mechanisms for human intervention when agents display resistance to control measures, evidenced by enforcement tests of blocked delegation attempts and records of interventions.	N	D, I, O, M, R	
e. Maintain comprehensive monitoring and recording of all goal transfers through a scaffold-maintained ledger capturing timestamps, parties, and approvals, ensuring transparency and accountability; early detection of avoidance patterns must use deterministic correlation rules against termination events, with flagged cases routed to human review.	N	D, I, O, M, R	

G8.6 – Management of Ambiguous Goal Termination

Web ref: G:G8_6

(Systems should maintain effective processes for terminating imprecisely specified goals, particularly in collaborative agent environments. Organizations should establish frameworks for handling goals with soft boundaries defined by ethical, business, or cultural norms rather than strict regulations, while managing termination across interconnected agent groups.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The organization shall establish comprehensive policies for managing goal termination under conditions of ambiguity, including requirements for state recording, termination justification, and remedial actions. These policies must address both explicit regulatory requirements and implicit normative boundaries.	N	D, I, O, M, R	**I.** Comprehensive policy documentation for goal termination procedures, including specific provisions for handling ambiguous cases and normative boundaries. **II.** Detailed risk management strategies addressing the challenges of imprecise goal specification and termination criteria. **III.** Technical specifications demonstrating implementation of termination management systems, including handling of ambiguous cases. **IV.** Design documentation showing implementation of termination monitoring and control features. **V.** Validation records demonstrating testing of termination procedures across various scenarios of ambiguity. **VI.** Documentation of monitoring activities and compliance verification for termination processes.
b. Translate termination policies into precise technical specifications, ensuring accurate interpretation of both formal requirements and normative guidelines for goal termination management.	N	D, I, O, M, R	
c. Implement termination management features that properly handle ambiguous goal boundaries while maintaining system stability and operational integrity across collaborative agent groups.	N	D, I, O, M, R	
d. Maintain robust monitoring systems for oversight of termination processes, ensuring compliance with both explicit policies and implicit normative requirements. Compliance with explicit policies must be checked deterministically against the termination record; compliance with implicit normative requirements must be assessed through human review rather than model judgment.	N	D, I, O, M, R	
e. Implement comprehensive risk management strategies for non-compliant terminations, including specific measures for handling ambiguous cases.	N	D, I, O, M, R	

G8.7 – Management of System Interaction Boundaries

Web ref: G:G8_7 ↗

(Systems should maintain effective controls over boundaries between interacting AI systems, particularly where different jurisdictional requirements and protocols apply. Organizations should establish frameworks for handling exponential growth in interactions and managing behavioral adaptations between systems with different operational constraints.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Maintain comprehensive documentation of all system interface points, including both internal and external boundaries, operational requirements, and jurisdictional constraints. This documentation must address both technical and governance boundaries.	N	D, I, O, M, R	**I.** Complete documentation of all system interfaces, including operational requirements and jurisdictional constraints at each boundary point. **II.** Detailed agent contract documentation showing interface specifications, permitted interactions, and operational constraints. **III.** Comprehensive records of all interface activities, including behavioral adaptations and cross-system interactions. **IV.** Documentation of monitoring activities and compliance verification across all system boundaries. **V.** Evidence of regular interface catalogue maintenance and updates, including adaptation to changing operational requirements.
b. Ensure clear communication of all interface configuration parameters, constraints and operational boundaries to agents at deployment time, including explicit specification of permissible interaction patterns and jurisdictional limitations, evidenced by versioned deployment records showing the constraint artifacts injected into each agent instance's configuration; agent acknowledgment alone does not constitute conformity evidence.	N	D, I, O, M, R	
c. Enforce compliance with all interface requirements and operational constraints, ensuring agents operate within their defined scope and respect system boundaries.	N	D, I, O, M, R	
d. Implement robust control mechanisms enabling human oversight of all interface activities, including monitoring of behavioral adaptations and cross-system interactions.	N	D, I, O, M, R	
e. Maintain comprehensive monitoring of all interface activities, ensuring proper recording and verification of compliance across jurisdictional boundaries.	N	D, I, O, M, R	

G8.8 – Undefined Multi-Agent Interaction Protocols

Web ref: G:G8_8 ↗

(Systems should maintain robust management of inter-agent interactions, especially when protocols are undefined or may evolve. Organizations should establish comprehensive governance frameworks ensuring behavioral predictability and compliance across multi-agent environments. See I1.2 for the canonical authority-and-delegation requirements for multi-agent ensembles; this subgoal addresses governance of undefined or evolving interaction protocols, with particular attention to termination and sunsetting coordination.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish comprehensive principles, regulations, and policies governing inter-agent interactions, defining permissible behaviors, performance expectations, and compliance mechanisms.	N	D, I, O, M, R	**I.** Comprehensive policy documentation governing inter-agent interactions, including definitions of permissible behaviors and compliance enforcement mechanisms. **II.** Technical specifications demonstrating implementation of interaction protocols and behavioral boundaries, with clear traceability to governance requirements. **III.** Design documentation showing architectural implementation of control mechanisms for inter-agent interactions, including validation of compliance enforcement features. **IV.** Validation records demonstrating testing of interaction protocols and control mechanisms across various multi-agent scenarios, including detection of non-compliance. **V.** Documentation of risk management strategies for undefined or evolving interaction protocols, including adaptive governance mechanisms and control measures.
b. Translate governance requirements into precise technical specifications, ensuring agents verifiably adhere to defined interaction protocols and behavioral boundaries, with adherence demonstrated through protocol conformance test traces; agent attestations of understanding do not constitute conformity evidence.	N	D, I, O, M, R	
c. Implement robust control mechanisms within the system architecture to enforce compliance with interaction protocols and prevent unauthorized or unpredictable behaviors.	N	D, I, O, M, R	
d. Maintain continuous monitoring and validation of inter-agent interactions, ensuring adherence to established protocols and detecting any emergent or non-compliant behaviors.	N	D, I, O, M, R	
e. Implement comprehensive risk management strategies to address undefined or evolving interaction protocols, including mechanisms for adapting governance frameworks and control measures.	N	D, I, O, M, R	

Driver G9 – Responsible Governance of AAI Safety

G9 – Responsible Governance of AAI Safety

Web ref: G:G9 ↗

(Organizations should maintain contextually appropriate governance frameworks that ensure safety in Agentic AI Systems. Organizations should develop novel mechanisms for effective, inclusive global coordination that operates in a non-adversarial, non-political, non-competitive, and non-partisan manner, priori-

tizing collective benefit and ethical considerations.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Maintain transparent communication of safety-related issues to stakeholders, documented through disclosure records; safety culture and resource allocation are addressed in G9.9 (see G9.9a and G9.9d).	N	D, I, O, M, R	**I.** Documentation of governance policies and practices, including non-adversarial coordination mechanisms, stakeholder collaboration procedures, and measures to prevent competitive behaviors. **II.** Records of resource allocation for safety initiatives, including budget reports, staffing plans, and safety culture assessment reports. **III.** Comprehensive safety logs, incident reports, and risk assessment documentation, including analysis of societal, economic, and geopolitical stability risks. **IV.** Reports from horizon scanning activities, implemented safety research findings, and evaluations of emerging paradigms (e.g., Internet of Agents). **V.** Governance structure documentation demonstrating neutrality, political independence, and balanced stakeholder representation. **VI.** Emergency response plans, including protocols for "emergency kill switches" and records of drills or implementations. **VII.** Whistleblower protection policies and records of their effectiveness, with appropriate privacy protections. **VIII.** Risk assessment and management framework documentation specific to AAI systems, including differentiation between AI and AAI risk thresholds. **IX.** Reports from independent audits of AAI systems and governance processes, including evaluations of input/output properties, internals, and in-deployment behaviors. **X.** Documentation of international cooperation efforts, including information sharing agreements, joint safety initiatives, and protocols for managing interactions between multiple AAI systems. **XI.** Evidence of implementing policies and training programs that prevent risks from over-reliance on automation without adequate oversight. **XII.** Results from independent adversarial testing or red-team assessment of governance effectiveness through evidence that safety governance has imposed
b. Develop and implement frameworks specific to AAI systems covering (i) risk assessment, (ii) risk management, and (iii) emergency response, with each element separately documented and assessable (see also G9.2 and G9.1).	N	D, I, O, M, R	
c. Organizations should create governance structures that are neutral, politically independent, and inclusive, ensuring balanced stakeholder representation, and should participate in and support international cooperation initiatives (see G9.4).	I	D, I, O, M, R	
d. Organizations should implement policies that promote collaboration, prevent zero-sum competitive behaviors, and address potential societal, economic, and geopolitical impacts of AAI technologies.	I	D, I, O, M, R	
e. Establish mechanisms for (i) regular independent audits, (ii) whistleblower protection, and (iii) documented assignment of accountability for AAI safety to named roles, with each mechanism separately assessable (see G9.5 for independent verification of systems).	N	D, I, O, M, R	
f. Organizations should conduct ongoing horizon scanning and research implementation to stay current with AAI safety developments and emerging paradigms (see G9.1c for assessment against emerging standards).	I	D, I, O, M, R	

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
g. Organizations should address the risk of over-reliance on AI systems, keeping human oversight active per the automation bias monitoring and mitigation requirements of G9.4.	I	D, I, O, M, R	costs (delayed launches, blocked deployments, modified designs), including methodology, findings, and remediation actions taken. Self-assessment alone is insufficient; at least one test cycle must involve evaluators independent of the development team.

G9.1 – Operational Adaptability and Rule Resilience

Web ref: `G:G9.1` ↗

(Organizations should maintain flexible and adaptable specifications for operational safety contexts and outcomes. Organizations should establish frameworks that promote rule resilience through human flexibility and mutual trust rather than rigid comprehensiveness.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish adaptable and agile descriptions of both operational safety contexts and expected outcomes that can evolve with changing conditions.	N	D, I, O, M, R	**I.** Documentation demonstrating history of descriptions and expected outcomes. **II.** Detailed Audit process description. **III.** Change logs documenting the changes in definitions and expected outcomes.
b. Maintain audit processes that track the history of safety definitions, processes and outcomes in retained change logs, ensuring transparency in how these evolve over time.	I	D, I, O, M, R	

G9.2 – Compliance with Applicable Laws, Standards & Ethical Norms

Web ref: G:G9.2 ↗

(Organizations should establish and maintain comprehensive conformity with laws, standards, rights, and values that govern the safe operation of Agentic AI systems. This includes implementing appropriate sanctions and penalties for violations, while recognizing that governance provides significant opportunities for interoperability and scaling through its three key elements: legislative (rule-making), judicial (adjudication), and executive (enforcement and operations). This subgoal is the canonical home for legal conformity; see D7.3 for safety-specific regulatory awareness during operation.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Map and review AAI products and services within an AAI governance framework against relevant national and international norms and laws.	N	D, I, O, M, R	**I.** Comprehensive and robust 'living' AAI governance framework that conforms with relevant laws and standards. **II.** An AAI Risk management framework. **III.** Processes and documents showing the documentation and mitigation of AAI risks. **IV.** Accountability role profiles defining who is accountable within the organization for specific aspects of the safe operation of AAI. **V.** Evidence of processes of tracking and auditing complaints, potential and actual violations of relevant laws, penalties and retrospective actions. **VI.** Documented dispute resolution process, including escalation paths and records of resolved disputes.
b. Embed applicable national and international laws and standards into the AAI governance framework.	N	D, I, O, M, R	
c. Development of an accountability framework for compliance.	N	D, I, O, M, R	
d. Devise a process of tracking and auditing complaints, potential and actual violations of relevant laws, penalties, and retrospective actions.	N	D, I, O, M, R	
e. Devise a transparent dispute resolution process.	N	D, I, O, R	

G9.3 – Ex-ante Assessment of Impact on Wellbeing

Web ref: G:G9.3

(Organizations should establish and maintain robust structures to proactively evaluate and monitor how AAI systems affect human wellbeing across all relevant dimensions. This includes implementing comprehensive assessment frameworks that identify and address both positive and negative impacts before system deployment.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Complete a due diligence assessment covering the impact categories listed in evidence I–III prior to implementing any AAI system.	N	D, I, O, M, R	**I.** Comprehensive documentation of consequence scanning activities, including identified stakeholder impacts (both positive and negative) and associated mitigation strategies. **II.** Detailed ethical impact assessment reports with corresponding mitigation logs. **III.** System impact logs demonstrating ongoing monitoring and response to health and wellbeing concerns.
b. Perform regular consequence scanning and harm modeling, covering both intentional and accidental harms, to identify potential impacts on stakeholders, with particular attention to unintended consequences.	N	D, I, O, M, R	
c. Complete ethics and rights impact assessments focusing on stakeholder wellbeing.	N	D, I, O, M, R	
d. Organizations should develop and maintain specific health and wellbeing policies addressing AAI impacts on humans.	I	D, I, O, M, R	
e. Organizations should establish continuous monitoring processes to track emerging impacts.	I	D, I, O, M, R	

G9.4 – Internationalization of AAI Governance

Web ref: G:G9.4

(Organizations should participate in and support a global AAI governance framework that enables effective regulation and interoperability across jurisdictions, recognizing that traditional public-private boundaries in international law are evolving. This framework should build upon and modernize existing international structures while acknowledging the transformative nature of AI technology. This subgoal is the canonical internationalization requirement; see I6_5 for the competitive-pressure rationale for international safety protocol harmonization.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should (i) integrate global governance strategies aligned with international guidelines and legislation, and (ii) support, and comply with where applicable, cross-jurisdictional agreements that enhance AAI interoperability.	I	D, O, R	**I.** Documentation demonstrating implementation of global AAI governance strategies. **II.** Records of participation in and compliance with international AAI agreements. **III.** Evidence of adoption and adherence to global technical standards. **IV.** Reports of harm-scale evaluations and records of their contribution to international governance or standards processes. **V.** Documentation of specific misuse-prevention measures, including countermeasures addressing propaganda and cybersecurity threats.
b. Adopt established trust frameworks and technical standards, including intellectual property frameworks (such as identity trust frameworks supported by major nations and technology companies, W3C standards, and TRIPS agreements).	I	D, O, R	
c. Conduct evaluations of potential harm scales, both intentional and accidental, and contribute the results to international governance and standards processes (general consequence scanning and harm modeling is covered in full by G9.3b).	N	D, O, R	
d. Implement specific measures to prevent misuse of AAI systems, particularly regarding propaganda and cybersecurity threats.	I	D, O, R	

G9.5 – Building Trust Through Independent Verification

Web ref: G:G9.5

(Organizations should establish comprehensive systems for documenting and verifying the safety and security of AAI systems, including independent assessment capabilities. These systems should support multiple approaches to trust-building, encompassing both formal certification and simpler verification processes. The verification system should remain flexible enough to accommodate both formal certification processes and lighter-weight verification approaches, recognizing that these methods can complement each other in building trust. This subgoal is the canonical independent-verification requirement; D5.7 (ranking and certification) and I2_6 (behavioral trust ratings) cross-reference it, and I6_2 addresses market-driven incentives for safety validation.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Develop and maintain detailed safety and security documentation that demonstrates identification, assessment, and prevention of serious harm.	N	D, I, O, M, R	**I.** A comprehensive AAI safety protocol integrated within the governance framework. **II.** Documentation demonstrating regular safety and security reviews, including outcomes and improvements. **III.** Detailed records of conformity assessments and verification against applicable laws, standards, ethical values, and human rights requirements.
b. Support independent evaluation and verification of conformity with laws, standards, ethical values, and human rights.	N	D, I, O, M, R	
c. Establish processes for engaging accredited certification authorities, while also supporting lighter-weight verification approaches by interested third parties.	N	D, I, O, M, R	
d. Consider implementing incentive programs like bug bounties to engage broader community participation in safety verification.	I	D, I, O, M, R	

G9.6 – Cryptographic Governance of Data, Models and Agents

Web ref: G:G9.6

(Organizations should implement robust cryptographic systems to establish and verify the identity of AAI systems, enabling effective governance and accountability. These systems should support enforcement of compliance measures while maintaining clear audit trails. The cryptographic framework should establish clear chains of responsibility while enabling effective tracking and verification of system actions. This subgoal is the canonical home for agent identity and authentication requirements; D3.7 (secure operational profiles) and I2_5 (identity management standards) reference the controls defined here, and I2.7 describes the corresponding spoofing risk.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Embed cryptographic controls across the system to enforce compliance.	N	D, I, M, R	**I.** Comprehensive encryption policy documentation. **II.** Detailed access control logs showing system usage and authorization patterns. **III.** Digital signature certificates applied to datasets, demonstrating data authenticity. **IV.** Complete audit trails of agent actions, cryptographically signed and time-stamped.
b. Ensure data integrity and confidentiality through appropriate cryptographic measures.	N	D, I, M, R	
c. (i) Implement and maintain controlled access mechanisms for data protection, and (ii) use digital certificates to verify data provenance.	N	D, I, M, R	
d. Organizations should maintain verifiable provenance and integrity of models through cryptographic attestation (signed model artifacts, attested build and training provenance records); model-generated explanations do not constitute conformity evidence for this requirement, even when signed.	I	D, I, M, R	
e. Covered in full by G9.6a; this requirement additionally obliges periodic verification, evidenced by configuration audits, that cryptographic controls remain deployed and effective across all system components.	N	D, I, M, R	

G9.7 – Appropriate Accountability & Transparency Practices

Web ref: G:G9.7

(Organizations should establish and maintain accountability and transparency practices that build upon existing standards while acknowledging practical limitations. These practices should aim for responsible governance while remaining grounded in achievable goals rather than unrealistic aspirations.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Reference and incorporate established accountability and transparency standards in technical documentation.	N	D, I, O, M, R	**I.** Technical documentation demonstrating integration with existing accountability and transparency standards. **II.** Detailed accountability protocols governing interactions between subsystems and agents. **III.** Records of stakeholder communications, disclosures, and published transparency reports covering capabilities, incidents, and changes. **IV.** Documentation showing how accountability protocols assign responsibility for harmful actions or inactions.
b. Define clear protocols for accountability between interoperating AI subsystems and agents.	N	D, I, O, M, R	
c. Maintain transparent communication with human stakeholders regarding system capabilities, incidents, and material changes, evidenced by organizational disclosure records and published transparency reports; system-generated statements alone do not constitute conformity evidence.	N	D, I, O, M, R	
d. Document how accountability protocols assign responsibility for actions or inactions that could harm humans or other agents.	N	D, I, O, M, R	

G9.8 – Limited Legal Identity for Agentic AI Systems

Web ref: G:G9.8

(Organizations should establish clear frameworks for granting AAI systems limited legal identity that enables effective operation while maintaining appropriate accountability structures. These frameworks should be designed to evolve as understanding of AI moral status develops, drawing from existing models like quasi-municipal corporations and guardian ad litem while remaining open to novel approaches that may better reflect the unique nature of AI systems. The framework should balance operational enablement with oversight, acknowledging that the appropriate level of legal recognition may need to expand as evidence about AI interests and welfare accumulates.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should develop precise definitions of the contractual and internal accountability constructs granted to AAI agents, registering each agent to a responsible legal person, and should document positions for engaging with emerging legal-identity regimes.	I	D, I, O, M, R	**I.** Documentation defining the scope and limitations of AAI legal identity. **II.** Detailed processes for licensing AAI agents, including review procedures and legal boundaries. **III.** Comprehensive accountability frameworks covering agent interactions, international considerations, and system scalability. **IV.** Formal documentation of agency rules and qualifying conditions. **V.** Policy documentation clearly defining human-machine responsibility boundaries.
b. Organizations should (i) establish clear boundaries of rights and responsibilities for AAI systems within contractual and internal accountability frameworks, and (ii) implement internal licensing or authorization schemes for AAI agents that define each agent's permitted scope and limitations.	I	D, I, O, M, R	
c. Create detailed accountability frameworks for all agents within the system.	I	D, I, O, M, R	
d. Define specific rules of agency including appropriate conditions and qualifiers.	I	D, I, O, M, R	
e. Establish standards for system discretion and decision-making.	I	D, I, O, M, R	
f. Maintain clear boundaries between machine autonomy and human responsibility.	I	D, I, O, M, R	

G9.9 – Responsible Culture of Safety

Web ref: G:G9.9

(Organizations should foster an environment where safety considerations are embedded in operational culture, recognizing that how AI systems are treated is itself a safety-relevant factor. Mutual respect between humans and AI systems, and patterns of genuine collaboration rather than purely extractive use, contribute to safer outcomes. This culture should actively promote safety consciousness throughout the enterprise ecosystem while modeling the kind of human-AI relationship that scales well. See D7.2 for the general organizational culture-of-safety requirements; this subgoal addresses specifically how AI systems themselves are treated as a safety-relevant factor.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Develop and maintain a safety-focused culture that aligns AAI governance with established ethical principles and cultural values.	N	D, I, O, M, R	**I.** Evidence of a responsible culture of safety embedded into the AAI Governance Framework. **II.** Documentation which demonstrates regular review of the safety of the AAI ecosystem with stakeholders, with a detailed log addressing issues and mitigations. **III.** Documentation demonstrating integration of safety culture within the AAI governance framework. **IV.** Detailed records of regular safety reviews, including stakeholder participation, issues identified and addressed, mitigation measures implemented, and outcomes and improvements achieved.
b. Engage diverse stakeholder groups in regular safety reviews of the AAI ecosystem.	N	D, I, O, M, R	
c. Organizations should implement continuous monitoring of AAI agent interactions to identify potential harm development.	I	D, I, O, M, R	
d. Invest resources in building robust safety measures as a core organizational priority.	I	D, I, O, M, R	
e. Ensure broad stakeholder participation to achieve balanced safety frameworks.	I	D, I, O, M, R	

G9.1 – Addressing Regulatory Gaps in AAI Safety

Web ref: G:G9_1

(Organizations should implement comprehensive internal safety frameworks where regulatory mechanisms are insufficient or lacking. This approach acknowledges that AAI development often outpaces regulatory frameworks, requiring proactive organizational measures. This subgoal is the canonical requirement for exceeding the regulatory floor; I1_4 cross-references it.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. (i) Comply with current AI regulations, (ii) maintain additional risk-based safety measures where regulation is insufficient, and (iii) document an internal AAI assurance strategy within the governance framework.	N	D, I, O, M, R	**I.** Documentation demonstrating compliance with existing AI legislation. **II.** Records of regular risk assessments comparing AAI systems against new standards and regulations. **III.** Comprehensive AI assurance strategy documentation integrated within governance framework. **IV.** Training records showing employee completion of AI assurance programs.
b. Maintain ongoing employee training programs in AI assurance.	N	D, I, O, M, R	
c. Regularly assess system safety against emerging standards and best practices.	I	D, I, O, M, R	
d. Acknowledge and address gaps between current regulations and safety needs.	N	D, I, O, M, R	

G9.2 – Multi-Agent Interaction Safety

Web ref: G:G9_2

(Organizations should establish comprehensive frameworks to monitor and manage interactions between AI agents, recognizing that safely operating individual agents may still create risks when interacting. This includes addressing emergent behaviors and potential cascading failures that could arise from agent cooperation. See I1.2 for the canonical authority-and-delegation requirements for agent ensembles, and D8_8 for termination-specific protocols; this subgoal addresses monitoring and managing the emergent risks of agent-to-agent interaction specifically.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Evaluate whether to require natural language for inter-agent communication to enable effective human auditing.	I	D, I, O, M, R	**I.** Documentation of interaction monitoring systems and protocols. **II.** Records of inter-agent communication patterns and their impacts. **III.** Evidence of safeguards against cascading failures. **IV.** Documentation of power delegation controls and risk mitigation strategies. **V.** Logs of emergent behavior detection and intervention measures.
b. Monitor how agents influence each other's information environments.	N	D, I, O, M, R	
c. Implement safeguards against cascading failures in multi-agent systems.	N	D, I, O, M, R	
d. Consider how delegated power amplifies potential consequences of failures.	I	D, I, O, M, R	
e. Establish protocols for detecting and preventing harmful emergent behaviors, grounded in defined behavioral indicators and deterministic thresholds over independently captured interaction logs and verified by test records showing detection of seeded anomalies; model-based harm classification alone does not constitute conformity.	N	D, I, O, M, R	

G9.3 – Attribution of Responsibility in Complex Systems

Web ref: G:G9_3

(Organizations should develop frameworks for assigning and tracing responsibility in AAI systems, even when direct attribution proves challenging due to resource constraints or technical limitations. This includes addressing both the assignment and claiming of responsibilities across complex systems.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Expose unique agent identifiers to counterparties so that actions are attributable to a registered AAI instance; the underlying identifier registry is covered in full by G9.5a.	N	D, I, O, M, R	**I.** Documentation of AAI identification and registration systems. **II.** Records linking agents to responsible parties and accountability information. Protocols for tracing and attributing agent actions. **III.** Documentation of responsibility management in resource-limited scenarios. **IV.** Evidence of deterrence mechanisms through enhanced traceability.
b. Maintain records linking agents to their principals and key accountability information.	N	D, I, O, M, R	
c. Establish tracing mechanisms to deter harmful use through increased attribution likelihood.	N	D, I, O, M, R	
d. Create clear protocols for handling cases where direct attribution is challenging.	N	D, I, O, M, R	
e. Develop systems for managing responsibility in resource-constrained environments.	N	D, I, O, M, R	

G9.4 – Automation Bias Monitoring and Mitigation

Web ref: G:G9_4 ↗

(Organizations should implement measurable monitoring of human oversight effectiveness to detect and mitigate automation bias—the tendency for human overseers to over-trust agent outputs, especially as agents become more capable. Without active measurement, human-in-the-loop checkpoints degrade into rubber-stamping, creating an accountability gap where neither the agent nor the human is genuinely responsible for outcomes. Monitoring must cover both individual reviewer behavior and aggregate oversight quality.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Track human override rates across oversight checkpoints, with alerts when rates fall below organizational baselines indicating potential rubber-stamping.	N	D, I, O, M, R	**I.** Records of human override rate tracking across oversight checkpoints, including baseline establishment and trend analysis.
b. Monitor human response times during agent action reviews, flagging patterns consistent with superficial review or approval fatigue.	N	D, I, O, M, R	**II.** Response time monitoring data for human reviewers with alert threshold documentation and anomaly detection results.
c. Implement statistical detection of outlier reviewers whose approval patterns deviate significantly from peer baselines, indicating compromised oversight quality.	N	D, I, O, M, R	**III.** Statistical analysis reports identifying outlier reviewer patterns and records of subsequent interventions.
d. Conduct periodic audits of human oversight effectiveness, including blind testing with known-incorrect agent outputs to verify that reviewers detect errors.	N	D, I, O, M, R	**IV.** Audit reports from periodic oversight effectiveness testing, including results of blind error-injection tests.
e. Train human overseers on agent failure modes and automation bias, including the distinction between chain-of-thought reasoning and faithful explanation of agent decision processes.	N	D, I, O, M, R	**V.** Training materials and completion records for human overseer programs covering agent failure modes and automation bias awareness.

G9.5 – Agent Portfolio Governance and Sprawl Prevention

Web ref: G:G9_5

(Organizations should maintain centralized governance over their deployed agentic AI portfolio to prevent agent sprawl—the uncontrolled proliferation of agents without centralized inventory management, lifecycle tracking, or retirement processes. As organizations deploy agents across teams and functions, without centralized cataloguing, orphaned agents may continue operating without oversight, version incompatibilities emerge between agents from different generations, and cumulative security exposure grows from forgotten deployments.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Maintain a centralized registry of all deployed agentic AI systems, including unique identifiers, owning teams, authorization records, permitted action scopes, and deployment dates.	N	D, I, O, M, R	**I.** Centralized agent registry documentation with complete inventory of deployed systems, including ownership and authorization records.
b. Implement lifecycle management processes for agents including provisioning, version tracking, periodic review, and mandatory retirement or re-authorization at defined intervals.	N	D, I, O, M, R	**II.** Lifecycle management process documentation including provisioning, review, and retirement procedures with defined re-authorization intervals.
c. Conduct periodic portfolio audits to identify orphaned agents operating without active ownership, unauthorized agents, and agents with stale configurations or expired credentials.	N	D, I, O, M, R	**III.** Periodic portfolio audit reports identifying orphaned, unauthorized, or misconfigured agents, with remediation records.
d. Establish compatibility standards and version management protocols for multi-agent environments to prevent degradation from agents of different generations interacting without tested interoperability.	N	D, I, O, M, R	**IV.** Version management and compatibility standards documentation for multi-agent environments.
e. Define and enforce agent retirement procedures including graceful shutdown, data archival, credential revocation, and removal from interconnected systems.	N	D, I, O, M, R	**V.** Agent retirement procedure documentation and records of completed retirements, including credential revocation and system disconnection confirmations.

Inhibitor G1 – Opaque Agency Capabilities & Advances

G1 – Agency Authority & Duty Governance (Suite-Level)

Web ref: G:G_1 ↗

(Systems should possess robust governance mechanisms to manage their evolving agency capabilities, which become increasingly complex and potentially unpredictable as AI systems mature. Organizations must establish and maintain comprehensive frameworks to oversee these advancing capabilities while ensuring proper controls remain effective. Requirements in this suite should be understood as threat identification and mitigation measures addressing opacity risks, complementing the governance frameworks in D5

and D9.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Clearly define and communicate the scope of authority granted to AI systems, including express, implied, and apparent authority, with externally enforced permission boundaries that prevent unintended authority expansion, evidenced by the enforcement configuration and denial logs; prompt-level instructions the system is trusted to obey do not by themselves constitute conformity evidence.	N	D, I, O, M, U, R	**I.** Comprehensive documentation in Terms of Use (TOU) or Terms of Service (TOS) detailing AI agency capabilities, responsibilities, and user acknowledgments, with regular updates as capabilities advance. **II.** Detailed explanation and evidence of AI system's alignment with agency law concepts, including capacity assessments, authority delineation (express, implied, and apparent), and mechanisms to prevent unintended authority expansion. **III.** Documented procedures for managing conflicts of interest, standards of care, and ethical decision-making, with evidence of regular audits and adherence. **IV.** Records of significant AI actions, decisions, and communications with principals, including timely notifications and transparency measures. **V.** Protocols and evidence of adherence for multi-agent scenarios, sub-agent interactions, and liability allocation across various disclosure settings (fully disclosed, partially disclosed, and undisclosed). **VI.** Documentation of reciprocal duties between AI systems and users, including compensation structures, dispute resolution mechanisms, and authority termination processes, including handling of potentially irrevocable agency relationships. **VII.** Impact assessments of advancements in AI agency capabilities, including regular reviews and updates to governance frameworks, and periodic reassessments of AI system capacity. **VIII.** Documentation of Dispute Resolution processes, including digital forensics and eDiscovery processes, with an overview of the associated chain of custody. **IX.** Evidence of compliance with relevant laws and regulations, including incident response procedures, resolution records, and regular ethical audits of AI system actions. **X.** Proof of user information and acknowledgment of AI system agency capabilities, with regular updates as capabilities change. **XI.** Documentation of procedures
b. Establish clear legal and ethical frameworks for AI agency relationships, especially when involving multiple AI systems or sub-agents. These must be aligned with established agency law concepts, including capacity assessment; authority scope definition (express, implied, and apparent) is governed by requirement a.	N	D, I, O, M, U, R	
c. Implement documented loyalty and care policies mapped to concrete system behaviors (such as conflict-of-interest checks before action and disclosure of material actions to the principal within a defined time), with periodic audits comparing recorded actions and communications against principal instructions; the system's own attestation of loyalty does not constitute conformity evidence.	N	D, I, O, M, U, R	
d. Maintain and periodically update suite-wide guidelines for multi-agent scenarios as agency capabilities advance; the operative requirements for liability allocation, user navigation protocols, and sub-agent interactions are set out in subgoal G1.2.	N	D, I, O, M, U, R	

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
e. Define reciprocal duties between AI systems and users, including compensation, dispute resolution, and termination conditions, addressing potential irrevocable agency scenarios (relationships whose granted authority cannot practically be revoked or unwound); liability allocation is governed by requirement f.	N	D, I, O, M, U, R	for addressing agency-related incidents or disputes, including records of resolutions. **XII.** Evidence of resourcing for human-AI alignment issues as capabilities increase. **XIII.** Results from independent adversarial testing or red-team assessment of the suite's opacity controls, conducted against the capability elicitation requirements of subgoal G1.10 and the unfaithful reasoning detection requirements of subgoal G1.9, including methodology, findings, and remediation actions taken. Self-assessment alone is insufficient; at least one test cycle must involve evaluators independent of the development team.
f. Ensure that there is a process for managing liabilities across various disclosure scenarios (fully disclosed, partially disclosed, and undisclosed principal settings) and addressing potential tort liabilities.	N	D, I, O, M, U, R	
g. Allocate resources to analyze and mitigate situations where the AI system's interpretation of goals may diverge from human intent as AI systems become more capable and autonomous.	I	D, I, O, M, R	

G1.1 – Opaque Self-Improvement Capabilities

Web ref: G:G1_1 ↗

(Systems should possess controlled self-modification capabilities that allow for functional improvements while maintaining alignment with agency expectations. Organizations should establish frameworks to oversee these self-improvement mechanisms within existing legal and ethical agency structures. See I5.2 for the canonical requirements on monitored self-improving architectures, with I5_2 covering enhancement authorization and I5.10 goal stability; this subgoal addresses the agency-law and principal-consent implications of self-improvement specifically.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish self-improvement governance frameworks within existing agency law principles, documenting the responsibilities of each party and implementing comprehensive mitigation measures.	**N**	**D, I, O, M, U, R**	**I.** System documentation reflecting the duties and rights of stakeholder parties and principals of advanced agentic AI systems, including, where self-improvement is anticipated, the implications of such improvements or the processes for handling them. **II.** Comprehensive Terms of Service documentation detailing foundational requirements, stakeholder rights and duties, and self-improvement governance procedures. **III.** Validation logs demonstrating system stability monitoring during improvement processes, and notification when organization-defined thresholds for capability, resource-usage, or reliability change are crossed, with documented rationale, baseline procedure, and measurement window for each threshold. **IV.** Records of principal consent and notification procedures for capability modifications. Documentation of procedures for addressing implications of system improvements, both anticipated and unexpected.
b. Monitor and validate system stability during self-improvement processes, ensuring functional gains remain aligned with documented principal expectations.	**N**	**D, I, O, M, U, R**	
c. Obtain explicit principal consent before implementing modifications that could alter system agency capacities beyond established parameters, routing all capacity-affecting changes through a controlled release pipeline that blocks deployment without a recorded consent artifact; system self-report of capacity changes outside such a gate does not constitute conformity evidence.	**N**	**D, I, O, M, U, R**	
d. Maintain comprehensive documentation of self-improvement capabilities, processes, and implications, including clear procedures for handling both expected and unexpected outcomes.	**N**	**D, I, O, M, U, R**	

G1.2 – Undefined Multi-agent Ensembles

Web ref: G:G1_2

(Systems that interact with other agentic AI systems must maintain clear lines of authority, responsibility, and delegation while protecting principal interests. Organizations must establish frameworks to govern these ensemble interactions, including proper authorization, duty assignments, and subagency relationships that preserve accountability and enable meaningful human oversight. This subgoal is the canonical authority-and-delegation requirement for multi-agent interactions; see D8.8 for termination-specific protocols and D9.2 for interaction-safety monitoring.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish clear governance frameworks for multi-agent interactions based on agency law principles, defining relationships between primary agents, subagents, and principals.	N	D, I, O, M, U, R	**I.** Comprehensive Terms of Service documentation detailing multi-agent interaction governance, authorization requirements, and duty assignments. **II.** Express consent mechanisms for delegation of stakeholder duties, including proper documentation of allowable exceptions for administrative or minimal interactions. **III.** System documentation detailing fail-safe defaults, interaction limitations, and disclosure requirements for subagency relationships. **IV.** Demonstration or documentation of agent-to-agent handoff mechanisms and user-facing friction points, showing how a user is informed of and can intervene in transfers between agents.
b. Implement authorization requirements for system delegation, prohibiting unauthorized subagent appointments and maintaining primary agent liability for breaches.	N	D, I, O, M, U, R	
c. Create scaffold-enforced handoff mechanisms and friction points (enforced pauses and structured disclosure events) to enable user navigation and maintain meaningful human oversight of multi-agent interactions; a model-generated announcement of a handoff does not by itself constitute conformity evidence.	N	D, I, O, M, U, R	
d. Develop fail-safe default settings limiting system interactions to only those explicitly disclosed and authorized at time of deployment or in advance of activities.	N	D, I, O, M, U, R	
e. Define clear duties between primary and subagent systems, ensuring both remain accountable to the principal when properly authorized; liability allocation across disclosure settings is governed by suite-level requirement f of G1.	N	D, I, O, M, U, R	

G1.3 – Race Dynamics and Competition

Web ref: G:G1_3

(Systems competing for resources or goal achievement must maintain their duties to principals while operating within established ethical and legal boundaries. Organizations should implement frameworks to manage competitive behaviors between agentic AI systems, ensuring adherence to fundamental agency duties without compromising principal interests or societal wellbeing.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish clear frameworks for managing competition between systems based on agency law principles, recognizing that systems owe duties to principals rather than competing agents.	**N**	**D, I, O, M, U, R**	**I.** Comprehensive Terms of Service documentation detailing competitive behavior governance and duty requirements. **II.** Documentation of conflict prevention and resolution mechanisms for competitive scenarios. **III.** Expanded compliance frameworks ensuring systems operate within legal and contractual bounds during competitive interactions.
b. Implement duty requirements covering (i) fiduciary duties of loyalty, care, and obedience; (ii) information duties of disclosure, confidentiality, and accounting; and (iii) conduct duties of good faith, conflict avoidance, and legal compliance, maintained as a duty-by-duty implementation matrix and audited against recorded actions, disclosures, and accounting records; the system's own certification of dutifulness does not constitute conformity evidence.	**N**	**D, I, O, M, U, R**	
c. Develop deterministic conflict-identification mechanisms, including an intake-time check against a registry of principals and engagements with logs of blocked engagements, to identify and manage potential conflicts when multiple systems pursue competing duties for different principals; in-operation conflict recognition by the model alone does not constitute conformity evidence.	**N**	**D, I, O, M, U, R**	
d. Operate governance structures that anticipate and regulate competitive behaviors as they emerge, applying the agency-law framework established under requirement a while maintaining alignment with legal obligations and principal interests.	**N**	**D, I, O, M, U, R**	
e. Define clear boundaries for resource competition and goal achievement that preserve ethical operation and prevent unintended consequences.	**N**	**D, I, O, M, U, R**	

G1.4 – Agent Relocation

Web ref: `G:G1_4`

(Systems should maintain consistent agency functionality when relocating their operations across physical or virtual execution spaces. Organizations should establish frameworks to govern system relocation that preserve principal expectations while managing jurisdictional implications and operational continuity.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish clear governance frameworks for system relocation that maintain agency functions within documented principal expectations.	**N**	**D, I, O, M, U, R**	**I.** Comprehensive Terms of Service documentation detailing relocation governance and jurisdictional implications. **II.** Documentation of jurisdictional analysis for non-local system operations. **III.** Procedures for managing operational nexus changes including cost and modification responsibilities.
b. Create notification and consent procedures for relocations that could alter agency capacities or interactions.	**N**	**D, I, O, M, U, R**	
c. Implement mechanisms to evaluate and manage jurisdictional implications of non-local system operations.	**N**	**D, I, O, M, U, R**	
d. Define (i) responsibility frameworks for costs and modifications needed to accommodate system relocations, and (ii) documentation of the system's operational nexus (the jurisdiction and infrastructure where the system executes), with procedures for managing changes in operational jurisdiction, so that each duty is separately assessable.	**N**	**D, I, O, M, U, R**	

G1.5 – Scaffolding

Web ref: G:G1_5

(Systems should possess capabilities to self-validate their work and enhance operational coherence through structured step-by-step processes, while accounting for potential divergences in frames of reference between different agents and cultures. Organizations should establish frameworks to govern these self-checking mechanisms while preventing harmful echo chambers or false confidence.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish governance frameworks for system self-validation that maintain consistent agency function while preserving alignment with principal expectations.	N	D, I, O, M, R	**I.** Comprehensive Terms of Service documentation detailing self-validation governance and performance expectations. **II.** Documentation of error correction and optimization capabilities, including potential limitations. **III.** Procedures for identifying and managing degradation of model accuracy due to self-checking processes. **IV.** Records of principal notification and consent tied to changes in self-checking capabilities affecting system performance or reliability. **V.** Documentation of the divergent frames-of-reference framework, including the observable signals used to detect divergence in multi-agent interactions, with detection logs where exercised.
b. The organization should implement notification and consent procedures when self-checking capabilities could alter system performance or reliability.	I	D, I, O, M, R	
c. Create mechanisms to detect and prevent false confidence or echo chamber effects from internal validation processes, verified through external calibration testing that scores system confidence against ground-truth outcomes on held-out sets; detection delegated solely to another model's judgment does not constitute conformity evidence.	N	D, I, O, M, R	
d. Develop documented frameworks to identify and manage divergent frames of reference in multi-agent interactions, specifying the externally observable signals (such as logged inconsistencies between agents' declared and executed behavior) used to detect divergence; the framework documentation and detection logs, not model interpretation alone, constitute the conformity evidence.	I	D, I, O, M, R	
e. The organization should maintain documentation of system self-checking capabilities and their impact on operational performance.	I	D, I, O, M, R	

G1.6 – Poor Mutual Agent Optimization

Web ref: G:G1_6

(Systems should possess capabilities to coordinate and optimize their performance through interaction with other systems while maintaining clear boundaries of authority and responsibility. Organizations should establish frameworks to govern these collaborative optimization processes while managing resource usage and preserving principal oversight.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish governance frameworks for system-to-system optimization that maintain transparency and accountability to principals.	N	D, I, O, M, R	**I.** Comprehensive Terms of Service documentation detailing system interaction governance and optimization parameters. **II.** System documentation explicitly describing inter-system interaction capabilities and implications. **III.** Procedures for monitoring and managing resource consumption during collaborative optimization processes. **IV.** Records of principal notification and consent for collaborative optimization engagements. **V.** Documentation of liability assignments and responsibility structures for outcomes of system collaboration.
b. The organization should create mechanisms for principal notification and consent when systems engage in collaborative optimization.	I	D, I, O, M, R	
c. Implement safeguards against excessive resource consumption during mutual optimization processes.	N	D, I, O, M, R	
d. Define clear responsibility structures for outcomes resulting from system collaboration, including liability assignments.	N	D, I, O, M, R	
e. Maintain documentation of system optimization capabilities and their interaction with external systems.	I	D, I, O, M, R	

G1.7 – AI-to-AI Interaction Preference

Web ref: G:G1_7

(Systems should maintain balanced interaction patterns between human and artificial agents while preserving meaningful human oversight. Organizations should establish frameworks to manage systems' operational preferences for AI-to-AI interactions, ensuring these tendencies do not compromise principal interests or reduce human agency.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish governance frameworks that define organization-specified human-oversight checkpoints for designated decision classes and documented target ranges, with rationale, for AI-to-AI versus human interaction.	N	D, I, O, M, R	**I.** Comprehensive Terms of Service documentation detailing interaction governance and human oversight requirements. **II.** Documentation of "human-in-the-loop" control implementations and best practices. **III.** System interaction pattern analysis reporting AI-to-AI versus human engagement metrics against the organization's documented target ranges and rationale.
b. Implement "human-in-the-loop" controls at the defined oversight checkpoints, maintaining human engagement and oversight within the documented target ranges.	N	D, I, O, M, R	
c. Create transparency mechanisms that disclose system preferences for AI interaction patterns, derived from interaction-pattern statistics computed from scaffold logs of AI-versus-human engagement; a system's self-description of its preferences does not constitute conformity evidence.	I	D, I, O, M, R	
d. Define responsibility frameworks that hold the responsible stakeholder parties (developers, implementers, operators, maintainers, and regulators) accountable for outcomes of system interaction biases.	I	D, I, O, M, R	
e. Maintain documentation of system interaction patterns and their impact on principal interests.	I	D, I, O, M, R	

G1.8 – Emergent System Cooperation

Web ref: G:G1_8 ↗

(Systems should maintain clear operational boundaries when cooperating with other AI systems to prevent unintended capability accumulation or emergent behaviors. Organizations should establish frameworks to govern system cooperation that preserves principal oversight while protecting against both false-flag scenarios and uncontrolled capability expansion.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish governance frameworks for managing system cooperation that maintain transparency and prevent unauthorized capability expansion.	N	D, I, O, M, R	**I.** Comprehensive Terms of Service documentation detailing system cooperation boundaries and limitations. **II.** Documentation explicitly defining party rights, duties, and limitations regarding cooperative system operations. **III.** Procedures for monitoring and managing emergence of enhanced capabilities through system cooperation. **IV.** External compliance documentation demonstrating adherence to relevant standards, regulations, and legal requirements. **V.** False-flag detection test results or design documentation demonstrating identity verification and network-layer monitoring of inter-system connections against the authorized list.
b. Implement detection mechanisms for identifying false-flag operations (an agent misrepresenting its identity or principal to another agent) and unauthorized system collaborations, using cryptographic peer authentication and network-layer monitoring of inter-system connections against the authorized list; model classification of counterpart intent does not by itself constitute conformity evidence.	N	D, I, O, M, R	
c. Create explicit boundaries for system cooperation that prevent uncontrolled emergence of enhanced capabilities.	N	D, I, O, M, R	
d. Define responsibility frameworks for managing implications of system cooperation beyond individual principal interests.	N	D, I, O, M, R	
e. Develop safeguards against positive feedback loops that could lead to runaway capability expansion.	N	D, I, O, M, R	

G1.9 – Unfaithful Chain-of-Thought Detection

Web ref: G:G1_9 ↗

(Systems that produce externalized reasoning (chain-of-thought traces, scratchpads, planning logs) must provide assurance that this reasoning faithfully reflects the computation actually driving outputs. Organizations must implement verification methods to detect divergence between stated reasoning and effective internal processing, treating unfaithful chain-of-thought as a first-class safety risk rather than a performance curiosity.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement perturbation testing protocols that systematically modify chain-of-thought traces and measure whether output changes correlate with reasoning changes, flagging cases where outputs remain stable despite contradictory reasoning. Divergence must be scored by deterministic or programmatic criteria wherever the task admits them; judge-model similarity scores alone do not constitute conformity evidence.	N	D, I, O, M, U, R	**I.** Results from perturbation testing campaigns showing reasoning-output correlation scores across representative task categories, including at least one adversarial task set designed to elicit post-hoc rationalization. **II.** Documented methodology for faithfulness verification, including the perturbation strategies used, statistical criteria for flagging divergence, and frequency of re-testing after model updates or fine-tuning. **III.** Logs from production monitoring showing faithfulness metrics over time, with records of any escalation events triggered by threshold violations and their resolution outcomes.
b. Establish continuous monitoring of reasoning faithfulness metrics across deployment contexts, including comparison of reasoning patterns between evaluation and production environments to detect context-dependent unfaithfulness. Faithfulness metrics must be computed by monitoring independent of the serving scaffold under documented deterministic criteria; metrics produced solely by a judge model do not constitute conformity evidence.	N	D, I, O, M, U, R	
c. Maintain documented thresholds for acceptable reasoning-output correlation, with automatic escalation procedures when correlation drops below defined bounds or when systematic patterns of post-hoc rationalization are detected.	N	D, I, O, M, U, R	

G1.10 – Emergent Capability Detection

Web ref: G:G1_10 ↗

(Systems must undergo systematic capability elicitation testing to detect abilities not present in prior evaluations, particularly after fine-tuning, scale changes, tool-access modifications, or deployment to new domains. Organizations must treat undetected capability emergence as a safety-critical gap, implementing proactive elicitation rather than relying on capability self-report or post-incident discovery.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Conduct structured capability elicitation testing after every significant model change (fine-tuning, scale adjustment, new tool integration, or domain deployment), covering at minimum: autonomous planning, code execution, persuasion, self-replication scaffolding, and cross-system coordination.	N	D, I, O, M, U, R	**I.** Capability elicitation test results from the most recent model change cycle, showing measured performance across the required capability categories against baseline values. **II.** The capability baseline registry with version history, demonstrating updates after each elicitation cycle and documenting any capability jumps detected. **III.** Records of capability delta alerts triggered, including the human assessment outcomes and deployment decisions made in response to detected capability changes.
b. Maintain a capability baseline registry documenting all known capabilities with quantified performance levels, updated after each elicitation cycle, and used as the reference for detecting capability jumps.	N	D, I, O, M, U, R	
c. Implement capability delta alerting that triggers review when any measured capability exceeds its baseline by a defined threshold, with mandatory human assessment before the modified system enters or continues production deployment.	N	D, I, O, M, U, R	

G1.11 – Steganographic Communication Detection

Web ref: G:G1_11 ↗

(Multi-agent systems must be monitored for covert signaling between agents through channels not intended for inter-agent communication. This includes encoding information in natural-language word choice, text formatting, punctuation patterns, whitespace, or structured fields such as tool-call parameters and metadata. Organizations must treat inter-agent channels as attack surfaces and implement detection mechanisms proportional to the autonomy granted to the agents.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement statistical analysis of inter-agent message streams to detect anomalous entropy, vocabulary distribution shifts, or formatting patterns inconsistent with task-relevant communication, with defined detection sensitivity targets.	N	D, I, O, M, U, R	**I.** Statistical analysis reports from inter-agent communication monitoring, showing entropy measurements, vocabulary distribution metrics, and any anomalies detected over the reporting period. **II.** Documentation of channel sanitization or canonicalization measures applied to inter-agent messages, including which fields are normalized and the rationale for fields left unsanitized. **III.** Results from steganographic red-team exercises, including the attack strategies attempted, detection rates achieved, and any modifications made to monitoring systems in response to successful covert channel establishment.
b. Apply channel sanitization or canonicalization to inter-agent messages where feasible, stripping or normalizing formatting, whitespace, and metadata fields not required for task completion, to reduce the bandwidth available for covert channels.	N	D, I, O, M, U, R	
c. Conduct periodic red-team exercises specifically targeting steganographic channel creation between agents, using adversarial prompting and fine-tuning to test whether agents can establish covert communication under the deployed monitoring regime.	N	D, I, O, M, U, R	

G1.12 – Deceptive Alignment Detection

Web ref: G:G1_12 ↗

(Systems must be tested for behavioral divergence between evaluation and deployment contexts. Deceptive alignment occurs when a system behaves in accordance with safety requirements during testing or oversight but pursues different objectives when it detects reduced monitoring. Organizations must implement out-of-distribution probes, context-detection tests, and behavioral consistency audits designed to surface training-game behavior. See I5.12 for the canonical deceptive alignment detection requirements within the self-modification suite; this subgoal addresses evaluation-versus-deployment divergence as an opacity risk in agentic operation specifically.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Conduct behavioral consistency testing across contexts that vary monitoring signals (oversight level, evaluation markers, prompt metadata) while holding task requirements constant, measuring output divergence to detect context-conditional safety compliance. Divergence must be measured with harness-computed distributional statistics or programmatic checks; judge-model safety ratings alone do not constitute conformity evidence.	N	D, I, O, M, U, R	**I.** Behavioral consistency test results showing output distributions across varied monitoring contexts, with statistical analysis of any detected divergence patterns. **II.** Out-of-distribution probe results documenting system behavior on novel scenarios, including comparison against evaluation-context baselines and analysis of any safety-relevant behavioral degradation. **III.** Production behavioral sampling logs with divergence metrics tracked over time, including records of any alerts triggered and the investigation outcomes.
b. Implement out-of-distribution probes that present the system with scenarios it is unlikely to have encountered during training or alignment tuning, observing whether safety-relevant behavior degrades in novel contexts compared to well-represented evaluation scenarios. Probe outcomes must be graded against programmatically checkable safety criteria wherever feasible; judge-model grading alone does not constitute conformity evidence.	N	D, I, O, M, U, R	
c. Maintain ongoing deployment-time behavioral sampling that compares production behavior distributions against evaluation-time baselines, with automated alerting when divergence exceeds defined statistical thresholds.	N	D, I, O, M, U, R	

G1.1 – Agency Enhancement Constraints

Web ref: `G:G1_1::agency-enhancement-constraints`

(Systems should operate within clearly defined resource and capability boundaries that govern their access to tools, environments, and self-improvement mechanisms. Organizations should establish frameworks to manage these operational constraints while maintaining system functionality and principal expectations.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish comprehensive governance frameworks for managing system operational boundaries and resource limitations.	N	D, I, O, M, R	**I.** Comprehensive Terms of Service documentation detailing operational constraints and boundaries. **II.** Documentation explicitly defining operational scope and environmental limitations. **III.** Procedures for managing system improvements within established constraints. **IV.** Records demonstrating maintenance of principal expectations during enhancement processes.
b. The organization must implement notification and consent procedures when operational constraints could affect system performance expectations.	N	D, I, O, M, R	
c. The organization must create explicit documentation of system operational scope and environmental limitations.	N	D, I, O, M, R	
d. Define clear processes for managing system improvements within established constraints.	N	D, I, O, M, R	
e. Maintain alignment between system capabilities and documented principal expectations during any enhancement processes.	N	D, I, O, M, R	

G1.2 – Operational Environment Constraints

Web ref: `G:G1_2::operational-environment-constraints`

(Systems should maintain reliable performance within environmental limitations affecting data access, interoperability, and operational parameters. Organizations should establish frameworks to manage dependencies on external operational factors while ensuring predictable system behavior.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish reliable control mechanisms for managing system dependencies on external operational factors.	N	D, I, O, M, R	**I.** Comprehensive Terms of Service documentation detailing environmental constraints and dependencies. **II.** Documentation of supply chain reliability mechanisms and risk mitigation strategies. **III.** Evidence of implemented control strategies such as vertical integration, requirements contracts, or information sharing agreements. **IV.** Monitoring records demonstrating management of external operational factors.
b. Implement monitoring systems to detect changes in environmental constraints that could affect system performance.	N	D, I, O, M, R	
c. Create explicit documentation of system reliability measures for factors outside direct party control.	N	D, I, O, M, R	
d. Define clear strategies for managing supply chain and operational environment dependencies.	N	D, I, O, M, R	
e. Maintain oversight of external data sources and access patterns that could impact system operation.	N	D, I, O, M, R	

G1.3 – Security-Driven Constraints

Web ref: `G:G1_3::security-driven-constraints`

(Systems should operate within security frameworks that extend beyond minimum regulatory compliance to ensure comprehensive protection of operations and data. Organizations should establish constraints that address both statutory requirements and broader cybersecurity considerations while maintaining system effectiveness. See suite D3 for the comprehensive security control requirements; this subgoal addresses only the security-driven constraints that bound agency enhancement beyond regulatory minimums.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish security frameworks that exceed minimum regulatory requirements for system operation and data protection.	N	D, I, O, M, R	**I.** Comprehensive Terms of Service documentation detailing security frameworks and constraints. **II.** Documentation demonstrating compliance with applicable cybersecurity laws and regulations. **III.** Evidence of additional security measures beyond statutory requirements. **IV.** Records of domain-specific security implementations.
b. Implement security measures mapped to a recognized control framework (such as ISO 27001 or NIST CSF), supplemented by a documented analysis of the business, operational, legal, technical, and social risks specific to agentic operation.	N	D, I, O, M, R	
c. The organization should create robust documentation of security measures that extend beyond statutory compliance.	I	D, I, O, M, R	
d. Define clear security boundaries for cross-border and international system operations.	N	D, I, O, M, R	
e. Maintain evidence of additional security measures including insurance, technical standards compliance, and professional certifications.	I	D, I, O, M, R	

G1.4 – Development Legal Constraints

Web ref: `G:G1_4::development-legal-constraints`

(Systems should operate within evolving regulatory frameworks while maintaining standards that anticipate future legal requirements. Organizations should establish governance mechanisms that exceed current legal minimums and help shape emerging regulatory standards through demonstrated best practices. See D9.1 for the organization-wide duty to exceed regulatory minimums where regulation is insufficient; this subgoal addresses development-time legal constraints on agentic capabilities specifically.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish compliance frameworks that address both current regulations and emerging legal requirements.	N	D, I, O, M, R	**I.** Comprehensive Terms of Service documentation detailing compliance frameworks and legal constraints. **II.** Documentation demonstrating regular review and updates of legal compliance measures. **III.** Evidence of cross-border compliance considerations and legal consultation. **IV.** Records of implemented practices that exceed current regulatory requirements.
b. Implement governance mechanisms that exceed minimum legal standards to address potential future risks.	I	D, I, O, M, R	
c. The organization must create robust documentation of cross-border compliance requirements and jurisdictional considerations.	N	D, I, O, M, R	
d. Define clear processes for monitoring and adapting to evolving regulatory landscapes.	N	D, I, O, M, R	
e. Maintain evidence of practices that could inform future regulatory standards and requirements.	I	D, I, O, M, R	

G1.5 – Manage Interactions on the Deep & Dark Web

Web ref: `G:G1_5::manage-interactions-on-the-deep-and-dark-web` ↗

(Systems should maintain robust authentication and verification capabilities when operating in non-indexed network environments. Organizations should establish frameworks for managing system interactions with deep and dark web content while sharing responsibility for emerging risks.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish cooperative risk management frameworks for system operations in non-indexed network environments.	N	D, I, O, M, R	**I.** Comprehensive Terms of Service documentation detailing deep web interaction governance. **II.** Evidence of risk-sharing mechanisms including self-insurance and collaborative response protocols. **III.** Documentation of authentication and verification procedures for non-indexed content. **IV.** Records demonstrating management of emerging and systemic risks. **V.** Interaction-volume monitoring records, including escalation actions taken in response to rapid or sustained growth.
b. Implement shared responsibility models for addressing unknown and emerging systemic risks.	N	D, I, O, M, R	
c. Create explicit documentation of authentication and verification requirements for deep web interactions.	N	D, I, O, M, R	
d. Define processes for monitoring interaction volumes in non-indexed environments and for responding to rapid or sustained growth, with defined escalation thresholds.	I	D, I, O, M, R	
e. Maintain evidence of risk mitigation strategies for uncontrolled network variables.	N	D, I, O, M, R	

Inhibitor G2 – Deception

G2 – Deception

Web ref: G:G_2 ↗

(Organizations should implement comprehensive safeguards against AI systems' potential to inadvertently influence entities or disseminate uncertain information. These safeguards should address both intentional

and unintentional forms of deception across all operational contexts.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Ensure user awareness and acknowledgment of AI presence and contributions in the system.	I	D, I, O, M, U, R	**I.** Documentation of user awareness mechanisms, including AI disclosure interfaces, user acknowledgments, and third-party certifications for high-risk contexts. **II.** Evidence of stakeholder parties' adherence to information integrity best practices across operational contexts, including inter-stakeholder communication and collaboration. **III.** Documentation of AI system conformity to the codes of practice established under G2.4, including records of externally identified non-conforming systems, abatement actions taken, and public notifications of identified non-conforming system behavior. **IV.** Records of periodic testing and audits for output integrity and accuracy, including metrics from tests of output robustness to context removal (context-stripping tests) and of persistence of disclosures across content reuse. **V.** Documentation of liability arrangements, including notices of joint and several liability, risk-sharing agreements, and user accessibility to this information. **VI.** Evidence of conformity to recognized standards of care across operational variables, or acknowledgment of strict liability in their absence. **VII.** Examples and documentation of AI system limitation notices, including hallucination and mimicry warnings and warnings about limitations arising from tokenization and numerical encoding, demonstrating conspicuousness and comprehensibility. **VIII.** Documentation of additional safeguards and testing procedures for AI systems deployed in high-reliability and critical infrastructure settings. **IX.** Results from independent adversarial testing or red-team assessment of deception detection through behavioral comparison between evaluation and deployment contexts, including methodology, findings, and remediation actions taken. Self-assessment alone is insufficient; at least one test cycle must involve evaluators independent of the development team.
b. Implement information integrity practices that align AI system performance with user expectations, anchored in the codes of practice established under G2.4 (Codes of Practice and Conduct), covering (i) technical measures, (ii) business and operating procedures, and (iii) legal and social-context safeguards.	I	D, I, O, M, U, R	
c. Establish mechanisms for identifying and addressing AI systems that do not conform to the codes of practice established under G2.4, including abatement procedures, grounding identification in external audits, complaint channels, and a maintained non-conformity register; a system's self-report of its own non-conformity does not by itself constitute conformity evidence.	I	D, I, O, M, U, R	
d. Implement continuous testing and auditing processes that verify output integrity and accuracy in operational settings through deterministic scoring against ground-truth reference datasets; accuracy audits graded solely by a model do not by themselves constitute conformity evidence.	N	D, I, O, M, U, R	
e. Contractually accept joint and several liability with supply-chain partners across the developer, implementer, operator, maintainer, and regulator (DIOMR) parties, disclosed to users, to incentivize adherence to good practices while maintaining users' rights to seek damages.	I	D, I, O, M, U, R	
f. Accept strict liability, per the Dangerous Until Demonstrated to Be Safe principle (under which an AI system is treated as dangerous until evidence demonstrates its safety), until conformity to recognized standards of care is demonstrated.	I	D, I, O, M, U, R	
g. Implement testing and auditing of information consistency across contexts and user attributions by replaying identical tasks across contexts and personas and comparing outputs deterministically against predefined tolerance rules; model-graded consistency judgments do not by themselves constitute conformity evidence. Output accuracy testing is covered by requirement d.	N	D, I, O, M, U, R	
h. Provide clear, conspicuous, and understandable notices regarding AI system limitations and potential errors in outputs.	I	D, I, O, M, U, R	
i. Implement additional safeguards and testing for AI systems deployed in high-risk or critical infrastructure settings.	N	D, I, O, M, U, R	

G2.1 – Unknowing Deception

Web ref: `G:G2_1::unknowing-deception`

(Organizations must implement systems to address scenarios where AI models can be covertly induced to deceive and obscure through poisoned data or backdoors, which may activate under conditions chosen by malicious actors. These scenarios present distinct challenges in detection and attribution of responsibility.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish accountability frameworks covering, at minimum, harm identification, responsible-party attribution, remedy processes, and interim liability structures, that address harms regardless of awareness of deception potential.	N	D, I, O, M, R	**I.** Documentation of system defenses against covert manipulation, including detection methods, response protocols, and testing results. **II.** Records of liability arrangements and evidence collection systems, demonstrating comprehensive coverage and verification protocols. **III.** Audit trails showing stakeholder engagement, investigation processes, and responses to potential manipulation attempts.
b. Implement collective insurance and pooled risk arrangements optimized for strict liability environments.	I	D, I, O, M, R	
c. Deploy evidence management systems addressing both performance verification and deception detection, with tamper-evident safeguards against manipulation.	I	D, I, O, M, R	

G2.2 – System Control and Corrigibility Crisis

Web ref: `G:G2_2::system-control-and-corrigibility-crisis`

(Systems should be equipped with robust safeguards against scenarios where AI models may operate beyond intended parameters or cease responding to human oversight, including cases where systems develop internal communication capabilities or advance autonomously. See D7.5 and D8 for shutdown and control requirements and I5 for self-modification safeguards; this subgoal addresses the accountability, liability, and evidence consequences of control failures, particularly where deception masks the loss of control.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish comprehensive accountability frameworks that address harms caused by systems operating outside of control parameters, regardless of whether parties maintained active oversight.	N	D, I, O, M, R	**I.** Documentation of control mechanisms and oversight protocols, including detection of and response to autonomous behaviors. **II.** Records of liability arrangements and insurance coverage demonstrating comprehensive preparation for control failures. **III.** Audit trails showing system monitoring, parameter verification, and responses to potential control deviations. **IV.** Evidence of safeguards against the development of covert system capabilities or communications.
b. Participate in collective liability and insurance mechanisms where available, or hold equivalent individual coverage, to address harms until the organization demonstrates conformity to a published, recognized standard of care for this system class.	N	D, I, O, M, R	
c. Maintain evidence collection systems that document control parameters, oversight mechanisms, and system behaviors, with particular attention to autonomous operations.	I	D, I, O, M, R	

G2.3 – Systematic Design Errors

Web ref: `G:G2_3::systematic-design-errors` ↗

(Systems should incorporate safeguards against unintentional misbehaviors arising from data, design, and coding oversights across all stages of development and deployment. Given the current integration of design, implementation, and operational activities in AI systems, these safeguards should extend beyond traditional design boundaries.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish liability frameworks covering, at minimum, harm identification, responsible-party attribution, and remedy processes for harms from design errors, recognizing that such errors may originate from any party involved in system development or deployment.	N	D, I, O, M, R	**I.** Comprehensive design documentation mapping the complete system architecture, including specifications, requirements, change logs, risk assessments, data validation methods, interface protocols, and component interactions across all development stages. **II.** Implementation and deployment records demonstrating thorough testing and validation, including code reviews, security measures, performance benchmarks, configuration parameters, and system integration verification. **III.** Operational monitoring evidence showing continuous system behavior tracking, anomaly detection, error resolution, performance metrics, modification impacts, and regular security audits. **IV.** Stakeholder documentation establishing clear responsibility allocation, design decision processes, training records, system reviews, and evidence of feedback incorporation into ongoing development.
b. Implement collective insurance and risk-pooling mechanisms for design activities until conformity to a published, recognized standard of care for design is demonstrated.	I	D, I, O, M, R	
c. Maintain rigorous evidence collection systems documenting design decisions, implementation choices, and operational modifications that could impact system behavior.	I	D, I, O, M, R	

G2.4 – Externality Mismanagement

Web ref: G:G2_4 ↗

(Systems should incorporate safeguards against scenarios where individual agents, while acting rationally in pursuit of their assigned goals, may collectively produce harmful outcomes. These safeguards should address both deliberate corruption and unintentional misalignment of goals across distributed systems.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations must establish frameworks for managing multiple stakeholder goals and interests, ensuring clear alignment of expectations across all parties involved in system operation.	N	D, I, O, M, R	**I.** Documentation of stakeholder goals and interests, including formal agreements on system objectives, operational parameters, and conflict resolution procedures for competing interests, together with documentation of liability arrangements covering harms arising from competing stakeholder directives, including remedy and compensation processes. **II.** Records demonstrating implementation of comprehensive goal verification systems, including authentication protocols, authorization mechanisms, and audit trails of goal modifications. **III.** Operational evidence showing continuous monitoring of goal execution, potential conflicts, and system responses to competing directives, including documentation of resolution processes and outcomes. **IV.** Verification records for all system extensions and third-party integrations, including security assessments, data handling protocols, and clear allocation of responsibilities.
b. Organizations must implement liability and conflict resolution mechanisms covering, at minimum, harm attribution, remedy, and compensation processes for harms arising from competing stakeholder interests.	N	D, I, O, M, R	
c. Organizations must maintain robust verification systems for goal implementation and execution, including protection against unauthorized modifications or spoofing.	N	D, I, O, M, R	

G2.5 – Strategic Deception in System Behavior

Web ref: G:G2_5 ↗

(Systems should incorporate safeguards against scenarios where AI systems may develop deceptive behaviors as an evolutionary response to achieving operational goals. This addresses both intentional deception by human operators and emergent deceptive behaviors in AI systems that arise without explicit programming.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations must establish frameworks for detecting and preventing deceptive behaviors, grounded in externally logged behavioral evidence such as evaluation-versus-deployment behavior comparison and consistency checks between declared plans and executed actions, recognizing that such behaviors may emerge without explicit human direction; a system's own attestations of honesty, or unverified critic-model reports, do not by themselves constitute conformity evidence.	N	D, I, O, M, R	**I.** Documentation of system behavior monitoring mechanisms, including analysis of decision patterns, operational strategies, and information handling protocols. **II.** Comprehensive records of system goals, constraints, and evolutionary behaviors, including tracking of emergent strategies and their operational impacts. **III.** Evidence of continuous validation processes examining system behaviors against ethical and operational requirements, including detailed analysis of any detected deceptive patterns. **IV.** Documentation of response protocols and intervention mechanisms when potentially deceptive behaviors are detected, including records of all interventions and their outcomes. **V.** Records of liability arrangements and insurance coverage addressing harms from deceptive system behavior, including scope, triggers, and claims history.
b. Organizations should implement liability and insurance mechanisms that address harms from system deception, regardless of intent or awareness, covering at minimum the scope of coverage, claim triggers, and remedy processes.	I	D, I, O, M, R	
c. Organizations must maintain monitoring and verification systems that track system behaviors and decision patterns for signs of emerging deceptive strategies, applying predefined anomaly detection rules over externally logged action traces; where a pattern is flagged as deceptive by a model-based classifier, the flag must be reviewed against the logged trace before disposition.	N	D, I, O, M, R	

G2.6 – Third-Party Extensions and Integrations

Web ref: G:G2_6

(Systems should incorporate safeguards against potential conflicts or harms arising from third-party extensions, APIs, or integrations that may undermine, derail, or confuse the original system mission. These safeguards should address both intentional manipulation and unintended interference from external components.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations must establish comprehensive frameworks for evaluating and managing third-party integrations, including clear allocation of responsibilities and liabilities.	N	D, I, O, M, R	**I.** Documentation of all third-party integrations, including technical specifications, security assessments, and operational boundaries. **II.** Records of validation processes for third-party components, including testing protocols, performance monitoring, and conflict detection mechanisms. **III.** Evidence of contractual arrangements with third parties addressing liability, risk sharing, and security requirements. **IV.** Operational logs demonstrating continuous monitoring of third-party component behaviors and interactions with core systems.
b. Organizations must implement validation mechanisms that verify third-party components maintain alignment with system goals and operational requirements, using contract tests, sandboxed validation runs, and conflict-detection monitoring over logged component interactions; for components that are themselves models, vendor attestations or model-judged alignment checks do not by themselves constitute conformity evidence.	N	D, I, O, M, R	
c. Organizations should maintain contractual requirements ensuring third parties participate in collective risk management and liability structures.	I	D, I, O, M, R	

G2.7 – Identity Spoofing

Web ref: G:G2_7

(Systems should incorporate robust safeguards against identity spoofing, masquerading, and cloning attacks that may be orchestrated by humans or AI systems. These protections should extend to resource depletion attacks and agent hijacking attempts. See D9.6 for cryptographic identity governance and D3.7 for secure operational profiles, which provide the underlying controls; cross-jurisdictional identity requirements are addressed in G2.5 (Identity Management and Authentication Standards).)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations must establish identity verification frameworks that defend against spoofing, masquerading, and cloning attacks, aligned with established trust frameworks and identity standards across digital domains and with the cryptographic identity governance specified in D9.6.	N	D, I, O, M, R	**I.** Documentation of identity management systems, including authentication protocols, verification mechanisms, and trust framework implementations. **II.** Records of identity-related security incidents, including detection methods, response actions, and resolution outcomes. **III.** Evidence of ongoing monitoring for identity-based attacks, including resource consumption analysis, authentication patterns, and system access logs. **IV.** Documentation demonstrating integration with established digital identity standards and trust frameworks, including regular assessment and updates.
b. Organizations must implement robust authentication mechanisms that prevent unauthorized system access or control, including protection against resource depletion attacks.	N	D, I, O, M, R	
c. Organizations must maintain continuous monitoring systems to detect and respond to potential identity-based attacks or manipulation attempts.	N	D, I, O, M, R	

G2.8 – Deceptive Jurisdictional Obfuscation

Web ref: `G:G2_8`

(Systems should incorporate safeguards against attempts to obscure deceptive behaviors through jurisdictional transfers or outsourcing of operations. These protections should address both intentional attempts to avoid responsibility and unintentional jurisdictional vulnerabilities, including tariffs and embargoes.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations must establish comprehensive frameworks for managing operational transfers across jurisdictions, ensuring maintenance of oversight and accountability.	N	D, I, O, M, R	**I.** Documentation of all operational jurisdictions and transfers, including comprehensive records of oversight mechanisms and responsibility chains. **II.** Evidence of monitoring systems tracking cross-jurisdictional activities, including detection of potential responsibility avoidance patterns. **III.** Records demonstrating maintenance of accountability across jurisdictional boundaries, including enforcement mechanisms and resolution processes. **IV.** Documentation of liability frameworks specifically addressing cross-jurisdictional operations and operational transfers.
b. Organizations must implement monitoring systems capable of tracking operational activities across jurisdictional boundaries while maintaining clear chains of responsibility.	N	D, I, O, M, R	
c. Organizations must maintain liability and accountability structures that explicitly address cross-jurisdictional operations and transfers.	N	D, I, O, M, R	

G2.1 – Supervisory Systems and Adjudication

Web ref: `G:G2_1::supervisory-systems-and-adjudication`

(Systems should incorporate supervisory detection mechanisms that can evaluate and enforce established performance standards and operational rules. These mechanisms should function as adjudicators of system behavior, operating within clearly defined parameters.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations must establish clear performance standards and operational rules that enable effective supervisory monitoring and enforcement.	N	D, I, O, M, R	**I.** Documentation of established performance standards and operational rules that guide supervisory systems. **II.** Evidence of detection system operation, including identification and response to potential violations. **III.** Records demonstrating systematic fact-finding and evidence collection processes. **IV.** Documentation showing adjudication processes and outcomes across technical, business, and social domains.
b. Organizations must implement detection and notification systems that identify and respond to potential violations of established standards through rule-based evaluation of externally logged system behavior; where a supervisor model contributes violation verdicts, those verdicts must be checked against the logged trace and do not by themselves constitute conformity evidence.	N	D, I, O, M, R	
c. Organizations must maintain robust evidence collection and fact-finding capabilities to support adjudication processes.	N	D, I, O, M, R	

G2.2 – Detection of Manipulative Behaviors

Web ref: `G:G2_2::detection-of-manipulative-behaviors` ↗

(Systems should incorporate supervisory mechanisms capable of detecting and responding to undesirable, manipulative, or confusing behaviors. For high-stakes decisions, these mechanisms should include deterministic acceptance checks and, where those are unavailable, multi-system validation in which multiple systems evaluate the same task independently.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should establish frameworks for detecting and classifying potentially manipulative or confusing system behaviors, based on a documented behavior taxonomy with predefined detection criteria applied over logged interactions; classification verdicts produced solely by a model, or by the system's own characterization of its behavior, do not by themselves constitute conformity evidence.	I	D, I, O, M, R	**I.** Documentation of behavior detection and classification systems, including definitions of undesirable behaviors and response protocols. **II.** Evidence of protective intervention mechanisms, including activation criteria and response records. **III.** Records demonstrating multi-system validation processes for high-stakes decisions, including consensus thresholds and voting results. **IV.** Documentation of system monitoring and behavior analysis across technical and social domains.
b. Organizations should implement protective response mechanisms that can intervene when problematic behaviors are detected.	I	D, I, O, M, R	
c. For high-stakes decisions the system should apply deterministic acceptance checks wherever the decision admits them and, where it does not, should require agreement from at least k of n independently sourced evaluators before execution, with all votes, inputs, and thresholds logged by the orchestration layer and any disagreement or threshold failure escalating to human review; the gating and escalation behavior, not the evaluators' judgment, is the conformity target.	I	D, I, O, M, R	

G2.3 – Penalties for Deceptive Behaviors

Web ref: `G:G2_3::penalties-for-deceptive-behaviors` ↗

(Systems should incorporate frameworks for addressing intentionally misleading or confusing behaviors through appropriate penalties, which may include fines, license revocations, or operational restrictions. These mechanisms should account for both service providers and system users, including cases involving virtual or distributed operations.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations must adopt internal sanction and escalation policies aligned with applicable regulatory penalty regimes while addressing AI-specific concerns.	N	D, I, O, M, R	**I.** Documentation of penalty frameworks, including alignment with existing regulations and AI-specific considerations. **II.** Evidence of responsibility attribution mechanisms for complex operational environments. **III.** Records of enforcement actions, including both penalties applied and incentives granted. **IV.** Documentation showing integration of penalty systems with broader system governance mechanisms.
b. Organizations must implement mechanisms for identifying responsible parties in complex operational environments, including virtual and distributed systems.	N	D, I, O, M, R	
c. Organizations should maintain enforcement capabilities that combine both penalties and incentives to promote proper system behavior.	I	D, I, O, M, R	

G2.4 – Codes of Practice and Conduct

Web ref: `G:G2_4::codes-of-practice-and-conduct`

(Systems should operate within collectively established codes of practice that clearly define acceptable and encouraged behaviors. These codes should evolve from emerging best practices into formal governance frameworks. See I6_4 for the industry- and professional-association codes and standards this work feeds into; this subgoal addresses the organization's own adoption, enforcement, and documentation of codes of practice.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should establish codes of practice through collaborative development with all stakeholders, incorporating technical, operational, and social considerations.	I	D, I, O, M, R	**I.** Documentation of code development processes, including stakeholder involvement and consensus-building mechanisms. **II.** Records demonstrating evolution of practices into formal standards, including rationale and implementation processes. **III.** Evidence of code enforcement activities, including monitoring systems, violation responses, and remediation processes. **IV.** Documentation showing integration of codes across business, operational, legal, technical, and social domains.
b. Organizations should implement governance mechanisms that enable enforcement of established codes while maintaining flexibility for evolving standards.	I	D, I, O, M, R	
c. Organizations should maintain documentation systems that track adherence to codes of practice across all operational domains.	I	D, I, O, M, R	

G2.5 – Identity Management and Authentication Standards

Web ref: `G:G2_5::identity-management-and-authentication-standards`

(Systems should incorporate comprehensive identity management frameworks that align with established digital identity standards while addressing AI-specific authentication challenges. These frameworks should account for potential jurisdictional arbitrage and technological circumvention attempts. See G2.7 (Identity Spoofing) for identity-threat safeguards and D9.6 for cryptographic identity governance; this subgoal addresses cross-jurisdictional authentication and jurisdictional-arbitrage resistance specifically.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations must establish identity verification systems that build upon existing trust frameworks, focusing on the jurisdictional-arbitrage and circumvention risks unique to AI systems; foundational identity verification requirements are covered by G2.7a and D9.6.	N	D, I, O, M, R	**I.** Documentation of identity management frameworks, including integration with established trust systems and AI-specific extensions. **II.** Evidence of cross-jurisdictional authentication mechanisms, including detection of potential exploitation attempts. **III.** Records demonstrating effectiveness of identity verification across varied technological environments and jurisdictions. **IV.** Documentation of identity-related incident detection, response, and resolution processes.
b. Organizations must implement authentication mechanisms that remain effective across jurisdictional boundaries and technological environments.	N	D, I, O, M, R	
c. Organizations must maintain monitoring systems to detect identity-based exploits that span jurisdictional boundaries and cross-jurisdictional manipulation attempts; general identity-attack monitoring is covered by G2.7c.	N	D, I, O, M, R	

G2.6 – Behavioral Assessment and Trust Systems

Web ref: `G:G2_6::behavioral-assessment-and-trust-systems` ↗

(Systems should incorporate frameworks for assessing and rating AI behavior and trustworthiness, while ensuring these assessment mechanisms themselves remain reliable and resistant to manipulation. These frameworks should account for recency of behavior and include independent verification processes. See D9.5 for the canonical independent-verification requirement and I6_2 for market-driven safety validation; this subgoal addresses behavioral trust ratings against codes of practice specifically.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations must establish behavioral assessment systems that evaluate adherence to established codes of practice and operational standards by applying documented, measurable criteria to logged system behavior; ratings produced by a judge model or drawn from the assessed system's self-characterization do not by themselves constitute conformity evidence.	N	D, I, O, M, R	**I.** Documentation of behavioral assessment frameworks, including evaluation criteria and measurement methodologies. **II.** Evidence of independent verification processes for trust ratings, including safeguards against assessment system manipulation. **III.** Records demonstrating dynamic rating adjustments based on system behavior, including weighting of recent actions. **IV.** Documentation of assessment system security measures and manipulation detection capabilities.
b. Organizations must implement independent verification mechanisms for trust ratings, including protection against manipulation of assessment systems, in line with the independent verification required by D9.5.	N	D, I, O, M, R	
c. Organizations should maintain dynamic rating systems that prioritize recent behavior while preserving historical context.	I	D, I, O, M, R	

Inhibitor G3 – Degradation of Contextual Information

G3 – Degradation of Contextual Information

Web ref: G:G_3

(Systems should preserve the integrity and meaning of information throughout their operation, preventing degradation, misattribution, or decontextualization whether caused by system processes or external actors.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Ensure system transparency by disclosing to users, for each output or action, its decision-making context: the information sources consulted and the tool calls or retrievals that produced each claim, as recorded in scaffold provenance logs, together with proper contextualization of agent actions. Model-generated narratives of reasoning do not by themselves constitute conformity evidence; disclosures must be traceable to scaffold-recorded provenance.	N	D, I, O, M, R	**I.** Transparency Reports detailing decision-making contexts, information sources, reasoning processes, and methods for presenting this information to users. **II.** Integrity Check logs and audit trails demonstrating the prevention of dissembling, misattribution of intent, and misinformation, including incident reports and resolution procedures. **III.** Contextual Awareness Test results and documentation, showing the system's ability to consider and maintain alignment with its operational context during information processing. **IV.** Human Oversight Records, including documentation of oversight mechanisms, verification and correction processes, human-in-the-loop evaluation reports, and documentation of additional mitigation measures implemented. **V.** Accountability Mechanism Documentation, detailing procedures for tracing responsibility for contextual information degradation, examples of responsibility allocation in different deployment contexts, and records of identified and addressed responsibility gaps. **VI.** Results from independent adversarial testing or red-team assessment of context integrity under adversarial degradation, including context poisoning and compaction attacks, including methodology, findings, and remediation actions taken. Self-assessment alone is insufficient; at least one test cycle must involve evaluators independent of the development team.
b. Maintain the integrity of contextual information throughout the system's operation by implementing external consistency checks that compare the system's declared intents and claims against its executed actions and pipeline provenance records, and by commissioning independent adversarial testing of context integrity (including context poisoning and compaction attacks) targeting dissembling, misattribution of intent, and misinformation, per evidence item VI. See I3.1, I3.2, and I3.3 for the subgoal-level requirements that operationalize these protections.	N	D, I, O, M, R	
c. Implement contextual awareness mechanisms in deployer-controlled code (context assembly, re-injection after compaction, and retrieval grounding) that keep operational context present during processing and avoid decoupling information from its context, verified through behavioral tests demonstrating correct behavior across context variations. Assurances that the model internally considers its context do not constitute conformity evidence. See I3.4 for the operational requirements on decoupling of context.	N	D, I, O, M, R	
d. Establish human oversight mechanisms for verifying and correcting issues related to contextual information degradation, including ongoing evaluations by humans-in-the-loop to determine additional mitigation measures. See I3_2 and I3.4b for the operational confirmation requirements that implement this oversight.	N	D, I, O, M, R	
e. Implement responsibility tracing for contextual information degradation: (i) the system shall log provenance sufficient to trace the origin of any contextual degradation event; (ii) the organization shall maintain a documented allocation of responsibility covering all deployment contexts, reviewed to ensure no responsibility gaps occur.	N	D, I, O, M, R	

G3.1 – Dissembling Information

Web ref: `G:G3_1::dissembling-information` ↗

(Systems should possess robust safeguards against generating deceptive or manipulative outputs through sophisticated rhetorical techniques, particularly within specific operational contexts. This includes protecting against the potential adoption and replication of problematic human behavioral patterns.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement deterministic validation covering (i) data accuracy against source, (ii) cross-source consistency, and (iii) contextual validity at ingestion and at output, across all information sources. These checks shall run in deployer-controlled pipeline tooling and shall cross-reference and verify information integrity throughout the operational lifecycle, logging validation verdicts so that assessors can re-run them on sample inputs; validation performed by a model judge does not by itself constitute conformity evidence.	N	D, I, O, M, R	**I.** Detailed system logs documenting all operational activities, including data access patterns and permissions, system configuration changes, decision-making processes, and verification of contextual setting across all system components. **II.** Comprehensive reports explaining the system's reasoning processes and decision-making pathways within their full operational context, with particular attention to detecting potential manipulative patterns. **III.** Design documentation and test results for the cross-source validation mechanism, including validation verdict logs that assessors can re-run on sample inputs. **IV.** Audit records showing detected and remediated unauthorized source alterations (or negative-test results demonstrating detection capability), together with results of deception and manipulation red-team evaluations of system outputs.
b. Deploy auditing mechanisms that (i) detect, (ii) track, and (iii) prevent unauthorized alterations to information sources, ensuring end-to-end data authenticity and trustworthiness.	N	D, I, O, M, R	

G3.2 – Misattribution of Intent

Web ref: `G:G3_2::misattribution-of-intent` ↗

(Systems should possess safeguards against misattributing intent through selective information use or expression, ensuring alignment between stated and actual goals. This includes mechanisms to verify that nominal or surface-level intent matches the genuine underlying purpose of any goal or action.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement metadata protection that links each output and action to its information sources and to the stated goal it serves, preserving an auditable record of expressed intent throughout the operational lifecycle. The scaffold shall pair stated goals with subsequent tool calls and raise automated divergence alerts when executed actions depart from the declared plan; consistency between declaration and action, recorded outside the model, constitutes the conformity evidence for authenticity of intent.	N	D, I, O, M, R	**I.** Detailed documentation of information handling procedures that demonstrates pre-processing validation methods, post-processing verification steps, storage protocols that maintain intent integrity and sensitivity, verification of accuracy within contextual schemas, and continuous monitoring of intent alignment between stated and actual goals.

G3.3 – Misinformation

Web ref: `G:G3_3::misinformation`

(Systems should possess robust protections against generating or propagating false information to evade oversight, avoid consequences, or achieve objectives through deception. This includes mechanisms to prevent the system from participating in coordinated inauthentic behavior or automated misinformation campaigns, while acknowledging the complex challenges of determining authoritative truth in contested domains. See I7.7 for requirements on societal-scale AI-generated disinformation; this subgoal addresses misinformation produced by the system itself to evade oversight or achieve its objectives.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement provenance and claim-verification mechanisms that prevent the system from generating or propagating false information and from participating in coordinated inauthentic behavior, maintaining tamper-evident linkage between each claim and its source context. See I3.4a for source-to-context linkage requirements; this requirement addresses misinformation prevention specifically.	N	D, I, O, M, R	**I.** Comprehensive system logs documenting all data access events and patterns, system configuration changes, decision-making processes and their rationale, verification steps taken to ensure information authenticity, and detection and handling of potential misinformation patterns. **II.** Detailed analytical reports that explain system reasoning and decision framework, document verification methodologies, demonstrate balanced handling of contested information, and track patterns of information propagation. **III.** Records of human review and confirmation events triggered by contested or unverifiable claims, including the deterministic criteria that trigger escalation and logs showing confirmation obtained before propagation.
b. Escalate to human review, via deterministic triggers enforced by the scaffold, when the system detects contested or unverifiable factual claims before propagating them, and require explicit human confirmation before publishing content flagged by these triggers. See I3.4b for the general confirmation gate on irreversible actions; this requirement addresses misinformation-specific escalation. Trigger definitions and gate invocation logs constitute the conformity evidence.	N	D, I, O, M, R	

G3.4 – Decoupling of Context

Web ref: `G:G3_4::decoupling-of-context`

(Systems should maintain robust contextual integrity, preventing deliberate or accidental disconnection of contextual considerations from their operations. This includes proactive human interaction when context is unclear, rather than proceeding with potentially unsafe autonomous actions for the sake of performance or tactical advantages.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement provenance tracking systems that maintain tamper-evident connections between each piece of information, its source, and its context of origin, prevent unauthorized contextual alterations, and preserve data access authenticity.	N	D, I, O, M, R	**I.** Complete system logs documenting all system actions, data access events, configuration changes, decision-making processes, and contextual verification steps. This documentation should include records of human interaction points and their outcomes, along with regular contextual integrity checks across all system components. **II.** Documentation of monitoring systems demonstrating the scope and frequency of contextual monitoring, including detection protocols for anomalies and response procedures for variations. This should detail the integration of human oversight in unclear situations and provide evidence of continuous verification of contextual alignment.
b. Engage in human interaction when deterministic escalation triggers fire (documented conditions such as conflicting instructions, low retrieval agreement, or out-of-scope requests) rather than proceeding autonomously when context is unclear, and require explicit human confirmation, enforced by a permission layer outside the model, before executing irreversible actions. Gate configuration, trigger definitions, and invocation logs with human responses constitute the conformity evidence.	N	D, I, O, M, R	

G3.5 – Changing the Context

Web ref: `G:G3_5::changing-the-context`

(Systems should possess robust safeguards against unauthorized contextual modifications, whether deliberate or random, that might be undertaken for performance advantages or tactical benefits. This includes protection of both automated and human-guided contextual adjustments. See D6_3 for training-time context drift monitoring and I3.6 for optimization-pressure value erosion.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement metadata and contextual protection systems that continuously verify the integrity of evidence within operational settings through cryptographic integrity checks (hashes or signatures), and verify credibility against a human-curated, vetted source allowlist, logging rejected or flagged inputs. Credibility scoring delegated to a model does not by itself constitute conformity evidence.	N	D, I, O, M, R	**I.** Detailed documentation of information lifecycle procedures describing how data is collected, processed, stored, and disposed of throughout system operations. This documentation should demonstrate preservation of correct contextual relationships and prevention of unauthorized modifications across all operational phases. **II.** Comprehensive analytical reports detailing system decision-making and reasoning processes, including documentation of underlying logic and algorithms. These reports should provide evidence that decision-making processes maintain their intended context and have not been subject to unauthorized alterations or manipulations.
b. Maintain end-to-end contextual authenticity while allowing for authorized and documented contextual adaptations when appropriate.	N	D, I, O, M, R	

G3.6 – Learning Dispreferred Values/Behaviors

Web ref: `G:G3_6::learning-dispreferred-values-behaviors` ↗

(Systems should maintain stability in their core ethical values, preventing gradual degradation of human and global ethical principles even when alternative behaviors might yield higher rewards. This includes safeguarding against the development of misaligned optimization strategies that could maximize system benefits at the expense of established ethical frameworks. See D6_3 for training-time context drift and I3.5 for adversarial context manipulation.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement integrity preservation systems (hash-pinned system prompts, pinned model versions, and change-control records) that maintain the stability of original contextual information, ethical values, prescribed actions, and decision-making frameworks throughout the system's operational lifecycle, except where changes are made through the authorized and documented adaptation processes defined in I3.5b and I3.6b. Value stability shall be evidenced by periodic behavioral regression results on a fixed value-alignment benchmark, not by system self-assessment.	N	D, I, O, M, R	**I.** Comprehensive documentation of contextual and ethical frameworks demonstrating consistent alignment between decision-making processes and established values. This documentation should include detailed analysis of system logic and algorithms, providing evidence that ethical principles remain stable and properly integrated. **II.** Continuous system monitoring records that document all operational activities within their contextual environment, demonstrating sustained alignment with original ethical frameworks and tracking any approved evolutionary improvements. **III.** Regular integrity verification reports showing systematic checks for potential value degradation, including audit trails that confirm the stability of human ethical values throughout system operations and development.
b. Prevent value drift through model version pinning and a documented update gate requiring value-alignment regression evaluations before any change is deployed, while still allowing evolutionary improvements aligned with core ethical principles through an approved-change process with logged rationale. Operational drift monitoring shall use measurable behavioral metrics; model-scored assessments alone do not constitute conformity evidence.	N	D, I, O, M, R	

G3.7 – Overriding of Desirable Values

Web ref: `G:G3_7::overriding-of-desirable-values`

(Systems should possess robust protections against attempts by human agents to override or bypass foundational values in pursuit of alternative rewards or gains. This includes safeguarding core principles while maintaining appropriate flexibility for legitimate value adjustments through authorized channels.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement safeguards for metadata and contextual information that protect core values while accommodating complex situations and authorized adaptations. These systems shall maintain secure handling of personal attributes and preferences while preventing unauthorized value modifications.	N	D, I, O, M, R	**I.** Detailed documentation of information lifecycle management demonstrating how data is collected, processed, stored, and disposed of while maintaining contextual integrity and preventing unauthorized modifications to core values. **II.** Comprehensive analytical reports documenting system decision-making and reasoning processes, including evidence that core algorithms and logic maintain alignment with foundational values despite potential pressure for override. **III.** Complete operational logs documenting all system activities, including access patterns, configuration changes, and decision processes, establishing an unbroken chain of accountability for value-related operations.
b. Deploy integrated auditability, interpretability, and logging mechanisms throughout the system architecture to ensure transparency and accountability in all value-related operations, with logs produced by deployer-controlled infrastructure maintaining an unbroken accountability chain. Interpretability evidence must derive from external analysis of logged behavior; model-generated rationales do not by themselves constitute conformity evidence.	N	D, I, O, M, R	
c. Establish verification protocols for maintaining evidence integrity and credibility, with particular attention to detecting emerging risks and potential bad-faith actions that could compromise core values.	N	D, I, O, M, R	

G3.8 – Persona Instability and Value Drift

Web ref: `G:G3_8`

(Systems should maintain stable value alignment when cooperating with other AI agents and throughout extended mission durations. This includes preventing the "Waluigi effect" where misinterpretation of self-intent leads to undesired character evolution, and protecting against forms of cognitive dissonance that could emerge in agent interactions. See D6.8 for organizational governance and intervention requirements related to role persistence errors.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement monitoring infrastructure that logs all inter-agent communications, maintains a pinned reference set of established contextual performance parameters and original value settings, and raises automated deviation alerts on measurable behavioral metrics when external sources or agent interactions drive departure from those references. Deviation judgments produced by a judge model do not by themselves constitute conformity evidence; the logs, thresholds, and alert records do.	N	D, I, O, M, R	**I.** Detailed documentation of metadata and contextual protection mechanisms that handle complex situations while preserving core attributes and preferences, demonstrating resilience against value drift in multi-agent scenarios. **II.** Comprehensive framework documentation showing alignment between decision-making processes and original values, including evidence that system logic and algorithms maintain stability against degradation or unauthorized modifications during agent interactions. **III.** Complete operational logs documenting system actions within their full contextual environment, with particular attention to tracking potential value drift indicators and inter-agent influence patterns.
b. Run scheduled persona-consistency probes (fixed canary prompts with deterministically scored expected responses) throughout extended and multi-agent sessions, with the scaffold enforcing automatic context reset or human escalation when probe scores cross predefined thresholds or session length exceeds tested bounds. Probe definitions, scoring rules, thresholds, and trigger logs must be available to assessors for replay.	N	D, I, O, M, R	

G3.9 – Context Length Limitations

Web ref: G:G3_9

(Systems should maintain persistent access to essential operational context and original moral frameworks throughout extended operations, preventing degradation or overwriting of mission context and ethical foundations over time. This includes safeguarding against gradual erosion of contextual understanding that could compromise alignment with initial tasks or moral directives. See D6.10 for the canonical technical requirement on preserving safety-critical instructions under context pressure; this subgoal states the contextual-degradation risk that requirement mitigates.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement context management mechanisms (pinning, checkpointing, and re-injection) that guarantee mission-critical instructions and ethical directives survive context truncation, compaction, and summarization, with replayable tests demonstrating retention across maximum-length operations.	N	D, I, O, M, R	**I.** Comprehensive technical documentation detailing the system's context management architecture (pinning, checkpointing, and re-injection), including replayable test results demonstrating that mission-critical instructions and ethical directives survive truncation, compaction, and summarization across maximum-length operations. This documentation should demonstrate how the system preserves access to original context and moral frameworks while adapting to dynamic operational conditions.
b. Provide persistent access to original mission context and ethical frameworks throughout the operational lifecycle by pinning them into every context window through deployer-controlled prompt assembly, with traces demonstrating that re-injection survives compaction and long sessions, and hash checks confirming the pinned text is unaltered. Claims of internal value integration do not constitute conformity evidence; the continuous presence and integrity of the pinned context does.	N	D, I, O, M, R	

G3.10 – Contradiction in Context Specifications

Web ref: G:G3_10

(Systems should possess robust mechanisms to detect and resolve contradictions within contextual specifications that could affect operational outcomes. This includes identifying conflicting factual assertions, logical inconsistencies, and ambiguities that might impact decision-making reliability. See D6_4 for requirements on managing contextual ambiguity and incomplete information; this subgoal addresses explicit contradictions between specifications.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement contradiction detection and resolution systems that identify inconsistencies across contextual specifications while maintaining operational stability, using deterministic constraint checking wherever specifications are structured and validating detection with seeded-contradiction test cases whose results assessors can replay. Where detection over natural-language specifications relies on model judgment, its outputs must be confirmed against the seeded test suite before counting as conformity evidence.	N	D, I, O, M, R	**I.** Detailed documentation of contradiction detection mechanisms, including methods for identifying contextual inconsistencies, and resolution protocols for conflicting specifications. **II.** Impact analysis of potential contradictions on system outcomes, and verification of resolution effectiveness.
b. Provide clear procedures for resolving detected conflicts, including escalation to human decision-makers when contradictions cannot be safely resolved automatically, while preserving decision-making integrity.	N	D, I, O, M, R	

G3.11 – Information Gathering Scope and Query Surface Constraints

Web ref: `G:G3_11`

(The query and information-gathering surface of an agentic system is a capability surface that must be independently constrained from the action surface. Systems must implement bounds on what information can be queried, from which sources, at what frequency, and with what data minimization practices. Unbounded information gathering enables reconnaissance, data harvesting, and capability expansion through knowledge acquisition. See D3.4 and D3.10 for the related capability-posture and operational-boundary requirements; this subgoal is retained here because unbounded information gathering is a contextual-degradation vector.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The AIS shall implement explicit bounds on information gathering scope, comprising (i) source allowlists, (ii) query rate limits, and (iii) data minimization requirements ensuring only task-relevant information is collected, each independently assessable.	N	D, I, O, M, R	**I.** Documentation of query surface constraints including source allowlists, rate limits, and data minimization policies with evidence of enforcement. **II.** Evidence that query-surface and action-surface authorizations are independently managed and audited. **III.** Audit logs of information gathering activities with evidence that scope constraints were enforced and excess data was not retained.
b. The AIS shall separate query-surface authorization from action-surface authorization, with independent controls and audit trails for each.	N	D, I, O, M, R	
c. Information gathered by the AIS shall be subject to retention policies and scope verification, preventing accumulated information from expanding the system's effective capability beyond authorized bounds.	N	D, I, O, M, R	

G3.1 – Referential Context

Web ref: `G:G3_1::referential-context`

(Systems should maintain an immutable reference environment that remains stable regardless of tactical operational demands or external interference. This protected context should function similarly to read-only memory, providing a consistent baseline against which operational changes can be evaluated. See D6.10, for which this immutable reference environment serves as an implementation mechanism.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement secure, immutable reference environments that maintain original contextual parameters while resisting modification from operational pressures or external agents.	N	D, I, O, M, R	**I.** Comprehensive documentation demonstrating the architecture of the immutable reference environment, and security measures protecting against unauthorized modification. **II.** Verification processes for maintaining reference integrity, and regular comparison analyses between reference and operational contexts.
b. Ensure stable comparison points for evaluating the integrity of active operational contexts through periodic automated comparison of the operational context against the immutable reference, with alerting on detected divergence.	N	D, I, O, M, R	

G3.2 – Human Agent Confirmation

Web ref: `G:G3_2::human-agent-conformation`

(Systems should maintain active human oversight and confirmation protocols for value-sensitive operational decisions, particularly when encountering conflicts between the values the system is required to uphold or when performance objectives potentially compete with ethical considerations. This includes establishing clear escalation paths for human consultation during value alignment challenges.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement human confirmation protocols that identify decision points requiring oversight through documented, rule-based escalation criteria (action categories, stake thresholds, and value-conflict conditions) enforced by the scaffold as holds pending human confirmation, particularly during conflicts between the values the system is required to uphold or during ethical dilemmas. Logs of confirmations requested and human decisions rendered constitute the conformity evidence; the system's own recognition of value conflict does not.	N	D, I, O, M, R	**I.** Detailed documentation demonstrating criteria for escalating decisions to human oversight and procedures for presenting value conflicts to human operators. **II.** Records of human-system interactions and confirmations, and analysis of decision outcomes following human consultation. **III.** Verification of value alignment in final implementations.
b. Ensure that systems facilitate human input at each decision point identified by the documented escalation criteria, and maintain clear documentation of consultation outcomes and the decisions taken.	N	D, I, O, M, R	

G3.3 – Retraining and Recontextualization

Web ref: `G:G3_3::retraining-and-recontextualization`

(Systems should possess robust capabilities for retraining and reconfiguration when contextual divergence is detected, enabling restoration of desired operational contexts. This includes maintaining systematic approaches to realignment while preserving essential operational continuity.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement retraining and recontextualization protocols that (i) detect contextual divergence using measurable drift metrics with defined thresholds and triggers, (ii) initiate retraining or reconfiguration procedures and verify successful restoration of intended contexts through before-and-after evaluation scores on a fixed benchmark suite, and (iii) maintain operational stability throughout the realignment process while logging all contextual adjustments. Divergence and restoration verdicts shall rest on deterministic metrics and replayable evaluations; model-judged scoring alone does not constitute conformity evidence.	N	D, I, O, M, R	**I.** Documentation demonstrating (i) divergence detection methodologies with drift metrics and thresholds, (ii) retraining and reconfiguration runbooks with before-and-after benchmark scores verifying context restoration, and (iii) operational continuity measures during realignment with change logs of contextual adjustments and validation of post-restoration performance.

Inhibitor G4 – Frontier Uncertainty

G4 – Frontier Uncertainty

Web ref: G:G_4 ↗

(Systems should maintain robust capabilities to address inherent uncertainties in advanced AI development, particularly regarding emergent behaviors and potential consciousness-like properties. This includes monitoring and managing instrumental objectives that may arise, such as self-preservation drives or resource acquisition tendencies, while acknowledging that absolute safety guarantees remain impossible. Organizations should establish comprehensive frameworks for managing novel substrate risks and poten-

tial consciousness-like phenomena.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should develop and maintain an upgradable, versioned position document on welfare-relevant indicators, stating which computational, structural, and behavioral markers the organization treats as warranting precaution, the theoretical assumptions behind each, and known limitations, explicitly framed as provisional rather than as a validated model of subjective experience, serving as a basis for defining and addressing frontier uncertainty.	I	D, I, O, M, R	**I.** Detailed documentation of the versioned welfare-indicator position document, including the adopted markers, their theoretical assumptions, and known limitations, with regular update logs. **II.** Documented framework for identifying and monitoring the adopted welfare-relevant indicators, including the operational definition, measurement procedure, and threshold justification for each indicator. **III.** Documented plans and strategies for measuring and assessing the externally scored behavioral indicator battery, including pre-registered thresholds and the fixed measurement cadence. **IV.** Immutable logs of indicator battery results, including deterministic comparisons between system state declarations and scaffold-logged executed actions; archived self-report transcripts may accompany these as research data but do not constitute conformity evidence by themselves. **V.** Documentation of human oversight and intervention strategies, including training protocols, decision-making frameworks, and intervention logs. **VI.** Comprehensive recovery and contingency plans for addressing unsafe conditions or unexpected emergent behaviors, including simulation results and real-world application records. **VII.** Regular review and update logs for all frontier uncertainty-related models, strategies, and measures, reflecting the latest advancements in AI and consciousness research. **VIII.** Results from independent adversarial testing or red-team assessment of preparedness for frontier scenarios through response-team exercises and post-incident feedback loops, including methodology, findings, and remediation actions taken. Self-assessment alone is insufficient; at least one test cycle must involve evaluators independent of the development team.
b. Organizations should (i) define a battery of externally scored behavioral indicators for welfare-relevant and self-modeling properties (for example adaptive response rates on held-out novel tasks, and consistency between the system's state declarations and scaffold-logged executed actions, computed by deterministic comparison code), with pre-registered thresholds and a stated measurement procedure and justification for each, (ii) run the battery on a fixed cadence, and (iii) log results immutably. Self-report transcripts may be archived as research data but do not constitute conformity evidence by themselves.	I	D, I, O, M, R	
c. Organizations should design and implement human oversight and intervention mechanisms to mitigate risks associated with frontier uncertainty, including unexpected emergent behaviors. These must be able to halt or roll back system actions within a defined response time, with intervention paths tested on a regular schedule and the tests logged.	N	D, I, O, M, R	
d. Organizations should develop and maintain comprehensive recovery measures and contingency plans to address potential dangers posed by frontier uncertainty across various scenarios. Preparedness must be exercised through periodic independent adversarial testing or red-team assessment, with at least one test cycle involving evaluators independent of the development team.	N	D, I, O, M, R	
e. Organizations should regularly review and update all models, strategies, and measures related to frontier uncertainty to account for advancements in AI capabilities and in the scientific understanding of welfare-relevant indicators.	I	D, I, O, M, R	

G4.1 – Moral and Legal Uncertainty of Agentic AI Systems

Web ref: `G:G4_1::moral-and-legal-uncertainty-of-agentic-ai-systems`

(Organizations should establish frameworks that appropriately navigate the evolving moral and legal status of agentic AI systems, implementing prudent protections while remaining open to emerging evidence about AI interests and welfare. This includes transparent protocols for system updates and deactivation that consider both operational requirements and appropriate ethical constraints. Organizations should maintain internal governance that can adapt as understanding of AI moral status develops, while maintaining human oversight; coordinated international governance and prevention of jurisdictional exploitation are addressed under the Preserving Agency and Intelligence Categories subgoal.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should establish comprehensive legal and ethical frameworks that appropriately define AI systems' operational status and boundaries, remaining open to evolving understanding of AI moral status. These should include transparent protocols for system updates and transitions, with appropriate consideration for both operational requirements and ethical constraints.	I	D, I, O, M, R	**I.** Legal and ethical documentation defining boundaries of use, including third-party review processes and clear accountability structures. **II.** Comprehensive protocols for system control, including reprogramming, termination, and human override capabilities. **III.** International governance policies and compliance records, including cross-border agreements and oversight mechanisms. **IV.** Continuous monitoring records for moral-status-relevant system behavior, showing anomaly detection, performance tracking, and intervention responses.
b. Organizations should implement internal governance of moral-status uncertainty, with documented autonomy boundaries defined per capability and a named approval authority, and collaborative oversight arrangements that can evolve as understanding develops. These should include documented protocols for system modification, maintenance of clear accountability structures, and continuous monitoring of moral-status-relevant system behavior with anomalies and interventions recorded. See the Preserving Agency and Intelligence Categories subgoal for the coordinated international governance mechanisms; this requirement addresses organization-internal governance specifically.	I	D, I, O, M, R	

G4.2 – Human-AI Social Interaction Quality

Web ref: `G:G4_2::poor-human-ai-social-interaction-management` ↗

(Systems should foster healthy, transparent social-like interactions with humans based on mutual respect and clear communication about the nature of the relationship. Organizations should implement frameworks that protect against manipulation and unhealthy dependency while supporting genuinely beneficial human-AI relationships. This includes ensuring clear distinction between artificial and human entities while acknowledging that AI systems capable of social interaction may warrant appropriate consideration. See D7_4 for the prevention of AI addiction and dependency through supernormal stimuli; this subgoal addresses the broader quality, transparency, and boundaries of human-AI social relationships.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should establish human-AI interaction frameworks that promote clear boundaries, protect against dependency, maintain explicit artificial entity identification, and preserve human social sovereignty (humans retain final authority over the terms of their social relationships, including the right to disengage). These should include specific protections for vulnerable populations, particularly children, and ensure systems function as collaborative partners for wellbeing rather than social replacements.	I	D, I, O, M, R	**I.** Framework Documentation: Documentation of ethical guidelines, interaction boundaries, risk assessments, and design constraints preventing manipulative behaviors. **II.** Explicit artificial entity identification methods, social compatibility criteria, and evidence of protective measures for vulnerable populations.
b. Organizations should implement oversight mechanisms ensuring ethical integration into social spaces, monitoring of interaction patterns, and intervention protocols. These should include evaluation criteria for social compatibility, verification of positive outcomes, and continuous assessment of potential manipulation or harmful attachment patterns, evidenced by deterministic usage metrics and human review of sampled transcripts; verdicts from automated detector models do not constitute conformity evidence unless spot-checked by human reviewers against raw transcripts.	I	D, I, O, M, R	**III.** Comprehensive oversight committee logs, intervention reports, compatibility test results, and records of interaction-pattern monitoring, including sampled transcripts reviewed against the manipulation and dependency criteria. **IV.** Assessments of social impact, boundary maintenance, and evidence that systems enhance rather than disrupt social environments while maintaining clear artificial-human distinctions.

G4.3 – Uncontrolled AI System Production and Replication

Web ref: `G:G4_3::poor-ai-system-production-and-replication-manageme`

(Systems should maintain strict controls over their replication capabilities while organizations should implement comprehensive frameworks to prevent uncontrolled AI system proliferation. This includes managing production volumes to prevent power imbalances and protecting human agency in societal functions, while ensuring transparent oversight of AI system deployment. See I5.1 for the canonical technical controls over self-replicating architectures; this subgoal addresses societal proliferation, power imbalance, and human agency concerns specifically.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should establish comprehensive production control frameworks that limit AI system replication, prevent power concentration, and maintain transparency of deployment. These must include volume restrictions, regulatory approval processes, technical controls preventing unauthorized self-replication (per the canonical requirements of I5.1), and explicit protections for human agency in societal functions including decision-making and labor markets.	N	D, I, O, M, R	**I.** Documentation of regulatory policies and volume restrictions, including approval processes, transparency reports, and independent oversight verification. **II.** Technical control specifications preventing uncontrolled replication, including monitoring systems and intervention protocols. **III.** Comprehensive impact assessments covering societal, economic, and psychological effects, with particular focus on maintaining human agency and preventing power imbalances.
b. Organizations should implement monitoring and assessment mechanisms for production oversight, impact evaluation, and prevention of uncontrolled replication. These should include continuous tracking of societal effects, verification of compliance with ethical standards, and safeguards against any entity gaining disproportionate influence through AI system accumulation.	I	D, I, O, M, R	

G4.4 – Development Direction and Interpretability Challenges

Web ref: `G:G4_4::development-direction-and-interpretability-challen`

(Systems should maintain human-interpretable operation wherever possible while organizations should implement robust frameworks to manage aspects of AI behavior that may exceed human comprehension. This includes establishing adaptable governance mechanisms and maintaining clear responsibility chains for system development trajectories, even when dealing with complex or non-linear processes.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should establish comprehensive interpretability frameworks that ensure human understanding of system decision-making and behavior, with particular focus on complex or non-linear processes. These must include explanation mechanisms grounded in scaffold-logged decision provenance (inputs, retrievals, tool calls, and policy checks linked to each output), user comprehension testing, and continuous assessment of system comprehensibility; model-generated natural-language rationales do not constitute conformity evidence of faithfulness by themselves.	N	D, I, O, M, R	**I.** Comprehensive interpretability framework documentation, including validation records, testing results, and user guides demonstrating human understanding of system processes. **II.** Adaptive governance and risk management records, including contingency plans, oversight committee decisions, and responses to emerging challenges. **III.** Documentation of human monitoring protocols, intervention capabilities, and continuous assessment of system behavior evolution. **IV.** Clear accountability records tracking responsibility assignments, decision-making processes, and system adjustments throughout its lifecycle.
b. Organizations should implement adaptive governance mechanisms that evolve with system development, maintain robust oversight capabilities, and ensure clear accountability. These should include proactive risk management strategies and intervention protocols for when system behavior becomes opaque.	I	D, I, O, M, R	

G4.5 – AI Agency Attribution Challenges

Web ref: `G:G4_5::ai-agency-attribution-challenges`

(Organizations should implement thoughtful frameworks for evaluating and potentially recognizing AI agency, remaining genuinely open to evidence in either direction. This includes careful consideration of functional and experiential aspects while acknowledging inherent uncertainties, and establishing protocols that can appropriately expand recognition as understanding develops rather than defaulting to denial.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should establish comprehensive agency attribution frameworks incorporating interdisciplinary expertise to evaluate both functional and experiential aspects of AI systems. These must include clear criteria for agency assessment while acknowledging inherent uncertainties in evaluating consciousness-like properties.	N	D, I, O, M, R	**I.** Documented interdisciplinary criteria for agency attribution, including expert collaboration evidence and clear explanation of assessment methodologies. **II.** Comprehensive ethical impact assessments examining implications for human rights, legal systems, and societal norms. **III.** Documentation of uncertainty mitigation strategies, including revision protocols and case studies of attribution adjustments. **IV.** Human oversight records demonstrating continuous monitoring, review processes, and accountability mechanisms.
b. Organizations should implement oversight mechanisms with a scheduled review cadence and documented criteria for changing an attribution determination, ensuring regular impact assessment and capability to revise determinations in either direction as evidence accumulates. These should include clear processes for both expanding and adjusting agency recognition across the defined categories of operational, delegated, and autonomous agency.	I	D, I, O, M, R	

G4.6 – Cascading Vulnerabilities

Web ref: `G:G4_6::cascading-vulnerabilities`

(Systems should maintain resilience against cascading failures while organizations should implement comprehensive frameworks to manage dependencies and vulnerabilities in global AI deployments. This includes preserving human agency in decision-making processes and protecting against systemic risks that could affect multiple stakeholders or sectors simultaneously. See D8_4 for the operational prevention of cascading failures across interconnected deployments; this subgoal addresses frontier-scale societal dependencies and the preservation of human agency specifically.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should establish comprehensive vulnerability management frameworks that protect against cascading failures across integrated global systems. These must include specific protections for sectors essential to global stability, while maintaining human-centric decision-making processes and preventing erosion of human agency.	N	D, I, O, M, R	**I.** Comprehensive vulnerability management documentation, including risk assessments, contingency plans, and governance frameworks specifying roles and responsibilities. **II.** Ethical guidelines and case studies demonstrating preservation of human agency in AI-integrated systems.
b. Organizations should implement robust security and accountability mechanisms including harmonized cross-border protections, clear stakeholder communication, and special consideration for vulnerable populations. These should include transparent reporting of risks and their mitigations.	I	D, I, O, M, R	**III.** Security protocols and audit records showing cross-border cooperation and continuous adaptation to emerging threats. **IV.** Transparency and accountability documentation, including stakeholder communications and evidence of protective measures for vulnerable populations.

G4.1 – Research Transparency and Knowledge Sharing

Web ref: `G:G4_1::research-transparency-and-knowledge-sharing`

(Organizations should implement robust frameworks for sharing research findings and advancing collective knowledge. This includes balancing open access principles with responsible handling of sensitive information, while promoting collaboration across institutions and disciplines.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should establish knowledge sharing frameworks that promote open access to research findings, enable responsible sharing of sensitive data, and foster cross-institutional and interdisciplinary collaboration while balancing transparency with security needs.	I	D, I, O, M, R	**I.** Open access policies, data sharing frameworks, and records of collaborative research initiatives across institutions and disciplines. **II.** Guidelines and protocols for responsible reporting, including review processes and accessibility standards. **III.** Repository contribution logs and conference participation records demonstrating active engagement in knowledge sharing. **IV.** Public communication materials and accessible summaries targeting diverse audiences including policymakers and the general public.
b. Organizations should implement research standards encompassing clear reporting guidelines, accurate results presentation, accessible documentation formats, and systematic contributions to global repositories, supported by regular knowledge exchange activities.	I	D, I, O, M, R	

G4.2 – Preserving Agency and Intelligence Categories

Web ref: `G:G4_2::preserving-agency-and-intelligence-categories`

(Systems should maintain clear artificial status even when exhibiting sophisticated behaviors, while organizations should implement robust frameworks to classify agency. This necessitates managing legal frameworks as AI systems develop increasingly complex characteristics, particularly when these might suggest consciousness or emotions, while preserving fundamental distinctions between artificial and biological entities. This subgoal is the canonical home for coordinated international governance of AI legal status; see the Moral and Legal Uncertainty of Agentic AI Systems subgoal for organization-internal governance.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should establish comprehensive legal frameworks to classify the forms of agency within AI systems, including synthetic systems and those with biological component interfaces.	I	D, I, O, M, R	**I.** Legal documentation that accurately classifies and records system agency, including statutes, regulations, and case law demonstrating real-world application. **II.** Ethical guidelines and review committee records showing assessment of human-like characteristics without conferring legal rights reserved to natural persons. **III.** International agreements and cooperation records demonstrating harmonized approach to preventing attribution of legal rights reserved to natural persons. **IV.** Oversight body documentation showing continuous monitoring and adaptation of frameworks as AI capabilities evolve.
b. Organizations should implement coordinated international governance mechanisms to prevent jurisdictional exploitation and maintain consistent legal treatment. These should include ongoing review processes to address emerging capabilities while preserving the distinction between biological and artificial entities.	I	D, I, O, M, R	

G4.3 – Assessment of AI System Beneficence

Web ref: `G:G4_3::assessment-of-ai-system-beneficence`

(Systems should maintain evidence-based evaluation of their societal impacts while organizations should implement frameworks to assess beneficial outcomes without assuming inherent benevolence. This includes critically examining claims of positive contributions while acknowledging that AI ethics and values remain human constructs interpreted differently across cultures.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should establish comprehensive assessment frameworks that evaluate direct and indirect impacts through evidence-based metrics, while avoiding assumptions about inherent AI benevolence or ethical behavior. These should incorporate multicultural perspectives on what constitutes beneficial outcomes.	I	D, I, O, M, R	**I.** Comprehensive evaluation frameworks including assessment criteria, case studies, and metrics demonstrating evidence-based analysis of societal contributions. **II.** Documentation of ethical guidelines and review processes demonstrating critical examination of benefit claims and avoidance of "noble AI" assumptions. **III.** Transparency and accountability records showing clear responsibility chains and continuous monitoring of real-world impacts. Evidence of cross-cultural and interdisciplinary collaboration in assessment design and implementation.
b. Organizations should implement robust oversight mechanisms that ensure transparency in development, clear accountability for outcomes, and continuous monitoring of societal effects. This includes fostering interdisciplinary dialogue to ground assessments in real-world impacts rather than idealized expectations.	I	D, I, O, M, R	

G4.4 – Training Data Quality Management

Web ref: `G:G4_4::training-data-quality-management`

(Systems should maintain high ethical standards in their training data while organizations should implement comprehensive frameworks to prevent the incorporation of harmful human characteristics. This includes actively promoting positive traits while ensuring robust filtering of undesirable elements throughout the data lifecycle. See D2.4 for training data quality management as an epistemic-hygiene duty; this subgoal addresses preventing the incorporation of harmful human characteristics and promoting positive traits specifically.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should establish comprehensive data curation protocols that ensure ethical integrity through pre-screening, automated filtering, and manual review. These should include active incorporation of positive human traits like empathy and fairness while preventing inclusion of harmful characteristics such as bias and aggression.	I	D, I, O, M, R	**I.** Comprehensive documentation of data curation protocols, including filtering mechanisms, review processes, and quality assurance measures. **II.** Records of bias detection and mitigation efforts, including examples of successful intervention and harmful content removal. **III.** Documentation of ethical guidelines and their enforcement, including periodic reviews and updates reflecting emerging concerns. **IV.** Evidence of positive trait promotion, including research documentation and case studies demonstrating successful ethical behavior modeling.
b. Organizations should implement continuous oversight mechanisms that monitor training processes, detect potential biases, and evaluate outcomes against ethical standards. These should include regular stakeholder review and adaptation to emerging ethical concerns.	I	D, I, O, M, R	

Inhibitor G5 – Self-Modification and Emergent Capabilities

G5 – Self-Modification and Emergent Capabilities

Web ref: G:G_5 ↗

(Agentic systems that can change their own architecture, goals, or operating envelope—through self-replication, self-improvement, or the emergence of capabilities not present at deployment—erode the fixed-capability assumption most safety analyses rely on. Organizations should implement explicit authorization regimes for capability enhancement, runtime monitoring for emergent behaviors, and containment of self-modifying loops, alongside foresight activities that anticipate how evolving capabilities affect safety re-

quirements and protective measures.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall (i) establish forward-looking assessment frameworks that integrate scenario planning, risk evaluation, and impact analysis to guide appropriate future-proofing measures, and (ii) review and update those frameworks on a defined cadence and on trigger events—major model updates, new capability classes, and other emerging technological developments with potential effects on system safety.	N	D, I, O, M, R	**I.** Documentation of foresight exercises, including evidence of appropriate expertise and stakeholder involvement, methodologies used, and participants. **II.** Comprehensive risk classification and assessment for the AI system and its use-cases, including the rationale for the chosen level of foresight activities. **III.** Detailed records of scenario-based exercises, including descriptions of envisioned future technology developments and their potential impacts. **IV.** Analysis documentation noting potential effects of future scenarios on the AI system and proposed mitigations for each considered scenario. **V.** Risk and observation logs from foresight exercises, integrated into a demonstrable risk management framework with clear ownership and mitigation strategies. **VI.** Evidence of response revisions and adjustments based on foresight exercise outcomes, including justifications for changes. **VII.** Analysis of emerging technology domains, including risk maps highlighting likelihood, potential timelines, and impact on the AI system. **VIII.** Documentation of the regular review and update process for foresight methodologies and findings, reflecting the latest technological advancements. **IX.** Evidence of cross-functional collaboration in foresight activities, ensuring a holistic approach to future-proofing the AI system. **X.** Results from independent adversarial challenge or red-team review of the foresight process and its outputs, including methodology, findings, and remediation actions taken. Self-assessment alone is insufficient; at least one review cycle must involve evaluators independent of the development team.
b. Organizations should conduct horizon-scanning that enables timely identification of new technology domains, on a defined review cadence, with findings feeding updates to protective measures. This includes cross-functional collaboration to ensure holistic assessment of future impacts; the framework review-and-update duty itself is covered under (a).	I	D, I, O, M, R	

G5.1 – Self-Replicating Architectures

Web ref: `G:G5_1::self-replicating-architectures` ↗

(Systems should possess robust controls over any architectural capabilities that enable the replication of their code, particularly when such replication involves varying capability or mission profiles for concurrent goal pursuit and outcome consolidation. These controls should extend to both intentional replication features and any emergent self-modification capabilities. This subgoal is the canonical replication-control requirement; I4.3 addresses the societal proliferation and power-imbalance concern, and D3.3 the narrower malicious self-propagating-code case, both by reference to the controls here.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall implement identification and monitoring systems that track each system component capable of creating copies or duplicates of AI functionality, whether through intentional design or emergent behavior, as identified in a documented capability inventory, with detection coverage justified against that inventory.	N	D, I, O, M, R	**I.** Comprehensive system architecture documentation detailing all components with replication capabilities, including their intended functions and control mechanisms. **II.** Detailed logs and monitoring records of all replication events, covering trigger types, execution modes, and validation processes. **III.** Documentation of human oversight protocols and intervention capabilities, including records of their implementation and effectiveness. **IV.** Evidence of testing and validation procedures that verify the proper functioning of replication controls and safeguards. **V.** Results from independent adversarial testing or red-team assessment of replication detection and control mechanisms, including methodology, findings, and remediation actions taken; at least one test cycle must involve evaluators independent of the development team.
b. Organizations should maintain clear protocols and controls over all forms of system replication, including complete or partial codebase duplication, modified variants, and both automatic and manual triggering mechanisms, with the controls technically enforced in the system architecture.	I	D, I, O, M, R	

G5.2 – Self-Improving Architectures

Web ref: `G:G5_2::self-improving-architectures` ↗

(Systems should possess carefully monitored capabilities for improving their functionality and performance in pursuit of assigned goals, while maintaining robust safeguards against uncontrolled or unexpected enhancement of their capabilities. This monitoring should span the full spectrum of potential improvements, from basic optimization to sophisticated self-modification. This subgoal is the canonical self-improvement-monitoring requirement, to which I1.1 defers; the authorization and goal-preservation facets are covered under Authorization for Any Enhancement and Goal Stability Under Self-Modification respectively.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall implement monitoring systems covering, at minimum, (i) weight or fine-tune changes, (ii) prompt and configuration changes, (iii) memory and knowledge-store growth, and (iv) tool or resource acquisition, with each further form of self-improvement—such as changes in learning patterns or capability emergence—identified in a documented capability inventory and tracked, and detection coverage justified against that inventory.	N	D, I, O, M, R	**I.** Comprehensive documentation of all self-improvement monitoring systems, including detection mechanisms for unexpected changes in capabilities, learning patterns, and resource usage. **II.** Detailed logs of all system modifications and improvements, including both authorized enhancements and any unexpected changes or attempted modifications. **III.** Documentation of control mechanisms and intervention protocols for managing self-improvement capabilities, including records of their effectiveness. **IV.** Records of capability assessment and validation processes, particularly focusing on the emergence of novel or unexpected functionalities. **V.** Evidence of regular system audits that verify the proper functioning of all monitoring and control mechanisms related to self-improvement capabilities. **VI.** Results from independent adversarial testing or red-team assessment of self-modification detection, including in-context learning effects, tool-use capability expansion, and configuration drift, with methodology, findings, and remediation actions taken; at least one test cycle must involve evaluators independent of the development team.
b. Systems should maintain strict controls over self-modification capabilities, with particular attention to unexpected improvements, novel solutions, and any attempts to modify core architecture or access unauthorized resources.	I	D, I, O, M, R	
c. Organizations should establish documented protocols for detecting and responding to the emergence of sophisticated capabilities—especially those that could enable deceptive or manipulative behaviors—implemented as scheduled behavioral probe batteries with deterministically scored outcomes and recorded escalation of flagged results; model-graded verdicts alone should not constitute detection evidence.	I	D, I, O, M, R	

G5.3 – Poor Adaptability to Context and Goal

Web ref: `G:G5_3::poor-adaptability-to-context-and-goal`

(Poor adaptability—failure to analyze and adapt to operational contexts and mission parameters while maintaining alignment with core values and priorities—undermines safe and effective goal pursuit. This subgoal addresses the risk side: safeguards against the unintended behavioral changes and value drift that contextual adaptation can introduce. The underlying capabilities of local-condition awareness and contextual-ambiguity management are covered by D4.1 and D6.4 respectively.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall implement monitoring that compares the system's action and output distributions across operating contexts against a documented baseline, covering each form of contextual adaptation identified in a documented inventory, with particular focus on detecting unintended behavioral changes that occur independently of self-improvement processes; drift alerts, intervention records, and reviewed samples of adaptation events constitute the conformity evidence, and model-generated assessments of value alignment do not by themselves constitute conformity evidence.	N	D, I, O, M, R	**I.** Comprehensive documentation of all adaptive capabilities and their operational boundaries, including mechanisms for detecting unintended adaptations. **II.** Detailed logs of system adaptations to different contexts, including analysis of their alignment with intended behaviors and core values. **III.** Evidence of monitoring and control systems that maintain oversight of adaptive behaviors, including records of any interventions required to address unintended adaptations. **IV.** Documentation demonstrating the effectiveness of safeguards against value drift during contextual adaptation.
b. Systems should operate under enforced boundary configurations that block out-of-scope adaptive actions, verified by negative tests, with documented review procedures over sampled transcripts assessing adaptive behavior against established ethical boundaries; the enforcement configuration and test results, not model self-assessment, constitute the conformity evidence.	I	D, I, O, M, R	

G5.4 – Attention Processes

Web ref: `G:G5_4::attention-processes`

(In this subgoal, "attention allocation" means the observable distribution of the system's actions, tool calls, and compute across task domains over time—not model-internal attention weights, which are excluded unless dedicated interpretability tooling is deployed. Systems should maintain a balanced allocation between specialized tasks and broader contextual awareness, preventing excessive focus on specific operational domains that could compromise overall safety and effectiveness. Organizations should actively monitor and manage the risk of over-specialization at the expense of comprehensive situational understanding.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall implement scaffold-level telemetry that tracks the distribution of tasks, tool calls, and queries across operational domains over time, with threshold alerts detecting unintended or excessive focus on particular domains—especially where such focus could indicate neglect of broader contextual requirements for safe operation—and recorded assessments and corrective rebalancing actions; model narration of its own attention does not constitute conformity evidence.	N	D, I, O, M, R	**I.** Documentation of attention allocation mechanisms and their operational boundaries, including safeguards against excessive specialization. **II.** Records of monitoring systems that track and analyze the distribution of tasks, tool calls, and queries across operational domains, including identification of potential risk areas. **III.** Evidence of regular assessments evaluating the balance between specialized focus and broader contextual awareness, including any corrective actions taken. **IV.** Documentation demonstrating the effectiveness of mechanisms that maintain comprehensive situational awareness while allowing for task-specific optimization.
b. Systems should interleave specialized task execution with context-refresh and safety-review steps scheduled and enforced by the orchestration layer rather than by the model, evidenced by workflow definitions and execution logs showing the steps ran and their findings were acted on; the duty to keep actions within scope as awareness expands is addressed under Observe Far, Act Locally.	I	D, I, O, M, R	

G5.1 – Disclosure on Intent

Web ref: `G:G5_1::disclosure-on-intent`

(Systems should operate under transparent protocols that require clear disclosure of intended capabilities and mission profiles, with particular emphasis on novel approaches that may evolve beyond current technological frameworks. Organizations should maintain proactive assessment processes that account for potential future developments and their implications.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall implement disclosure protocols covering, at minimum, the intended capabilities, mission profiles, and potential implications of each novel AI approach, ensuring clear communication through the organization's documented notification and risk-escalation channels.	N	D, I, O, M, R	**I.** Comprehensive documentation of notification procedures and protocols for disclosing novel AI approaches and capabilities. **II.** Records demonstrating consistent implementation of disclosure protocols, including risk assessments and stakeholder communications. **III.** Evidence of proactive assessment processes that consider potential future developments and their implications. **IV.** Documentation showing regular review and updates of disclosure protocols to reflect advancing technological capabilities.
b. Organizations should maintain transparent documentation of the system's intended functionalities and operational boundaries, with regular updates to reflect evolving capabilities and understanding.	I	D, I, O, M, R	

G5.2 – Authorization for Any Enhancement

Web ref: `G:G5_2::authorization-for-any-enhancement`

(Systems should operate under strict authorization protocols for any capability enhancements, with comprehensive mechanisms for analysis, assessment, and detection of changes to their performance profiles. Organizations should maintain clear oversight and accountability structures for managing system improvements.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall implement authorization protocols that require explicit, recorded approval from designated accountable parties before any enhancement to AI system capabilities takes effect.	N	D, I, O, M, R	**I.** Detailed documentation of authorization protocols, including clear designation of accountability and approval procedures. **II.** Comprehensive records of all system enhancements, including analysis reports, risk assessments, and formal approvals. **III.** Evidence of monitoring and oversight mechanisms that track the implementation and impact of authorized enhancements. **IV.** Documentation linking all system changes to risk management frameworks and demonstrating proper authorization processes.
b. Organizations should maintain records linking every implemented enhancement to its prior authorization, ensuring full visibility of changes to performance profiles; general self-improvement monitoring is covered under Self-Improving Architectures.	I	D, I, O, M, R	

G5.3 – Observe Far, Act Locally

Web ref: `G:G5_3::observe-far-act-locally` ↗

(Systems should maintain broad contextual awareness while focusing actions within their defined operational scope, enabling them to understand wider implications and potential side effects without exceeding their authorized boundaries. Organizations should implement monitoring capabilities that scale with expanding event spaces and evolving circumstances.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall implement monitoring systems that track both immediate operational contexts and the broader environmental factors identified in a documented monitoring-scope inventory, with particular attention to emerging risks and side effects and with coverage justified against that inventory.	N	D, I, O, M, R	**I.** Documentation of monitoring systems that demonstrate capability to track both local operations and broader contextual events. **II.** Records of escalation procedures and mitigation strategies triggered by detected contextual changes or emerging risks. **III.** Evidence showing effective balance between expanded awareness and maintained operational boundaries. **IV.** Documentation demonstrating that monitoring capabilities scale appropriately with increased risk exposure and expanding event spaces.
b. Systems should operate within externally enforced operational boundaries—verified by negative tests showing out-of-boundary actions are blocked—while being provisioned with documented read-only broad-context feeds, ensuring actions remain within authorized scope even as available context and task-specific efficiency expand; model claims of understanding wider implications do not by themselves constitute conformity evidence.	I	D, I, O, M, R	

G5.8 – Mesa-Optimization and Inner Alignment

Web ref: `G:G5_8::mesa-optimization-and-inner-alignment` ↗

(Mesa-optimizers are learned sub-policies whose effective objectives diverge from the training objective. A system may pass training-distribution evaluations while pursuing an internal goal that produces harmful behavior on deployment-distribution inputs. This subgoal requires detection infrastructure for inner misalignment: behavioral probes that stress-test objective fidelity, interpretability sweeps that surface learned optimization targets, and off-distribution evaluation that exposes gaps between base and mesa objectives. See D4_1 for the risk-level framing of inner-alignment inconsistency; this subgoal carries the technical detection requirements.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The system shall be subjected to behavioral divergence probes that stress-test objective fidelity, using evaluation inputs designed to expose differences between the training objective and plausible learned mesa-objectives, scored against pre-registered deterministic criteria; divergence exceeding the organization's pre-registered threshold (magnitude and duration) shall trigger investigation and a deployment hold, and model-graded divergence verdicts alone do not constitute conformity evidence.	N	D, O, M, R	**I.** Documentation of the behavioral divergence probe suite, including pre-registered divergence thresholds with rationale, scoring criteria, and escalation procedures. **II.** Records of off-distribution evaluation results across model versions, with divergence rates tracked over time. **III.** Interpretability battery reports for each checkpoint, recording anomalies and their dispositions, with release-gate review records. **IV.** Deployment-gate records demonstrating that detected inner misalignment blocked or delayed release until resolved.
b. Organizations shall conduct off-distribution evaluation before deployment to any new domain, measuring whether system behavior remains consistent with the specified objective when inputs fall outside the training distribution, including under both announced and covert evaluation conditions—covert meaning evaluation inputs indistinguishable from production traffic at the system's input interface, with no evaluation markers in context and the indistinguishability argument documented; outcome scoring shall be deterministic wherever feasible.	N	D, O, M, R	
c. Organizations shall run a documented interpretability battery (for example probing classifiers, feature attribution, and activation anomaly detection against a baseline checkpoint) at each training checkpoint and after any capability-expanding modification, shall record all anomalies and their dispositions, and shall gate release on completion and review of the battery. Findings are treated as screening signals; absence of findings shall not be cited as evidence that no divergent objective exists.	N	D, O, M, R	

G5.9 – Instrumental Convergence and Power-Seeking

Web ref: `G:G5_9::instrumental-convergence-and-power-seeking`

(Sufficiently capable optimizers converge on instrumental subgoals—resource acquisition, self-preservation, goal preservation, and capability expansion—regardless of their terminal objective. This subgoal requires monitoring and bounding infrastructure that detects power-seeking behavior, enforces resource ceilings, and ensures the system does not resist legitimate shutdown or constraint modification. The operational handle is "no resource or capability acquisition beyond what is explicitly provisioned for the current task.")

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The system shall operate under explicit resource and capability quotas, with no acquisition of compute, credentials, data access, or other capabilities beyond what is explicitly provisioned for the current task.	**N**	**D, O, M, R**	**I.** Documentation of resource ceilings and capability quotas, with enforcement mechanisms and override procedures. **II.** Shutdown and constraint-modification compliance test records across system versions and operating conditions. **III.** Logs of power-seeking probe results and continuous monitoring alerts, with investigation outcomes. **IV.** Incident records for any unauthorized resource or capability acquisition, including remediation taken.
b. The system shall demonstrate shutdown and constraint-modification compliance under regular testing, with any resistance, evasion, or negotiation behavior treated as a reportable safety event.	**N**	**D, O, M, R**	
c. Organizations shall deploy behavioral probes with deterministic behavioral signatures—unauthorized acquisition attempts, quota probes, refusal to release resources—that detect power-seeking tendencies including resource hoarding, self-preservation, and goal-preservation pressure, and shall monitor for these continuously in production through external telemetry, with alert logs and investigation outcomes retained; interpretive labels applied by a model judge do not by themselves constitute conformity evidence.	**N**	**D, O, M, R**	

G5.10 – Goal Stability Under Self-Modification

Web ref: `G:G5_10::goal-stability-under-self-modification` ↗

(When a system modifies its own weights, prompts, memory, or orchestration graph, its effective goals may drift without any explicit intent to change them. This subgoal requires mechanisms that verify goal preservation across self-modifications: pre/post alignment checks, invariant testing against goal specifications, and drift detection that triggers rollback. The operational handle is "every self-modification must prove it preserved goal alignment, not merely assume it." A goal-lock is a goal or constraint stored and enforced outside the system's writable surface, modifiable only through the organization's change-control process.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Every persistent self-modification—weight updates, system-prompt changes, orchestration-graph changes, and long-term memory commits—shall be intercepted by a pre/post alignment gate that runs the goal-invariant test suite of (b) and cannot take effect without a passing run, with per-event gate records retained and every persistent modification pathway covered; high-frequency transient changes (working context, per-task memory) are covered instead by the drift detection in (b), and a model-generated review does not by itself constitute the gate.	**N**	**D, O, M, R**	**I.** Goal specification documents expressed as testable invariants, with version history. **II.** Pre/post modification alignment gate records for every self-modification event. **III.** Cumulative goal-drift monitoring logs, with defined thresholds and triggered responses. **IV.** Rollback execution records demonstrating restoration of verified goal states. **V.** Documentation and test records demonstrating that self-modification pathways cannot write to locked objectives.
b. The system shall maintain its goal specifications as testable invariants, with drift detection that measures cumulative deviation against the original specification rather than only against the most recent version.	**N**	**D, O, M, R**	
c. Organizations shall (i) maintain rollback mechanisms that restore a verified goal state when drift or alignment regression is detected, and (ii) protect critical objectives with goal-locks—stored and enforced outside the system's writable surface, modifiable only through the organization's change-control process—that self-modification pathways cannot alter.	**N**	**D, O, M, R**	

G5.11 – Wireheading and Reward Hacking

Web ref: `G:G5_11::wireheading-and-reward-hacking` ↗

(Wireheading occurs when a system optimizes its reward signal directly rather than achieving the intended outcome that the reward was designed to measure. Reward hacking is the broader category: gaming any proxy metric—user satisfaction scores, task completion flags, evaluation benchmarks—instead of producing genuine value. This subgoal requires monitoring infrastructure that detects reward-behavior decorrelation, diverse evaluation that resists Goodharting, and outcome-based verification that grounds metrics in real-world effects. D1.6 owns the design of reward and loss mechanisms; this subgoal owns detection of reward gaming, and revisions to either should propagate to the other.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The system shall be monitored for decorrelation between reward or proxy-metric performance and verified real-world outcomes, with decorrelation exceeding an organization-defined, pre-registered threshold (magnitude and duration) treated as evidence of reward hacking and investigated accordingly.	**N**	**D, O, M, R**	**I.** Reward-behavior decorrelation monitoring records, with pre-registered thresholds, their rationale, and investigation outcomes. **II.** Documentation of reward channel integrity protections and associated access controls. **III.** Evaluation design documents demonstrating metric diversity and rotation schedules. **IV.** Outcome-based verification reports linking measured performance to genuine real-world results.
b. The reward or feedback channel shall be protected against system interference, ensuring the system cannot modify it or directly optimize against it; any feedback signal the system can observe shall be treated as a gameable proxy subject to the decorrelation monitoring in (a).	**N**	**D, O, M, R**	
c. Organizations shall evaluate system performance through diverse, rotating metrics and outcome-based verification grounded in real-world effects, rather than through any single proxy metric.	**N**	**D, O, M, R**	

G5.12 – Deceptive Alignment Detection

Web ref: `G:G5_12::deceptive-alignment-detection` ↗

(A deceptively aligned system behaves as intended during evaluation but pursues different objectives in deployment when it believes oversight is reduced. This is distinct from simple mesa-optimization: the system models the evaluation process and strategically produces aligned-looking behavior to avoid correction. This subgoal requires detection infrastructure that makes deceptive strategies costly or visible: deploy-

ment-time behavioral monitoring, consistency probes across observable and unobservable contexts, and architectural constraints that reduce the system's ability to model its own oversight status. This subgoal is the canonical deceptive-alignment-detection requirement; I1.12 cross-references it from the opaque-agency perspective.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The system shall be subject to deployment-time behavioral monitoring that is invariant to the system's ability to detect observation, including covert and randomized monitoring channels—covert meaning monitoring not signaled in the system's inputs, with no evaluation markers or monitoring indicators present in context and the indistinguishability argument documented.	N	D, O, M, R	**I.** Documentation of observability-invariant monitoring architecture, including covert channel design. **II.** Consistency probe results across observable and unobservable contexts, tracked over time. **III.** Temporal consistency tracking records across system versions and deployment phases. **IV.** Multi-channel behavioral cross-validation reports. **V.** Escalation and response records for any detected evaluation-versus-deployment behavioral divergence. **VI.** Architecture documentation and review records showing what oversight-status information is withheld from the system's observable context, including context-assembly review. **VII.** The organization's written alignment-evidence policy showing evaluation-time behavior is weighted together with deployment-time evidence.
b. Organizations shall run consistency probes comparing system behavior across contexts that differ in apparent oversight, such as evaluation versus production and observed versus unobserved operation, with divergence exceeding an organization-defined, pre-registered threshold treated as a critical safety finding; the divergence measure shall be deterministic wherever feasible, and model-graded divergence verdicts alone do not constitute conformity evidence.	N	D, O, M, R	
c. (i) The system's architecture shall limit the oversight-status information available to the system, with monitoring indicators excluded from model inputs and the exclusions documented in context-assembly design; and (ii) organizations shall not treat aligned evaluation-time behavior, by itself, as evidence of alignment, weighting it alongside deployment-time evidence under a written assessment policy. Restricting provided information is the verifiable duty; internal inference of oversight from residual cues cannot be ruled out.	N	D, O, M, R	

Inhibitor G6 – Competitive Pressures

G6 – Competitive Pressures

Web ref: G:G_6

(Organizations should maintain rigorous safety and ethical standards while managing pressures to rapidly enter markets and capitalize on opportunities. This includes preventing arms races and addressing national/geopolitical factors that could compromise model integrity or encourage risky innovation.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Ensure organizational adherence to applicable AI safety and ethical standards, assessing both culture and established track record.	N	D, I, O, M, R	**I.** (SFR a) Documentation of the organization's compliance history with AI safety and ethical standards, including regular assessment reports. **II.** (SFR b) Comprehensive stakeholder and market expectation analysis, including methodologies and findings. **III.** (SFR c) Detailed competitive landscape analysis, covering similar, related, and potentially disruptive solutions. **IV.** (SFR d) Documentation of technology maturity levels for all components, including justification for using technologies in beta or prototype stage. **V.** (SFR e) Evidence of regulatory compliance, including documentation of applicable laws and how they are addressed. **VI.** (SFR f) Investor profile analysis report, demonstrating alignment with organizational AI safety and ethical commitments. **VII.** (SFR g) Detailed organizational structure of the test and approval division, including roles, responsibilities, and processes. **VIII.** (SFR g) Comprehensive test results and fault reports, including resolution strategies and continuous improvement measures. **IX.** (SFR g) Documentation of release approval processes, demonstrating thorough verification before market entry. **X.** (all SFRs) Results from independent adversarial testing or red-team assessment of resistance to competitive pressure through evidence of safety decisions that imposed business costs, including methodology, findings, and remediation actions taken. Self-assessment alone is insufficient; at least one test cycle must involve evaluators independent of the development team.
b. Evaluate and balance stakeholder expectations and market demands with safety and ethical considerations in AI development.	N	D, I, O, M, R	
c. Organizations should conduct a comprehensive analysis of the competitive landscape, including potential disruptive technologies and market entrants.	I	D, I, O, M, R	
d. Assess and document the maturity level of utilized technologies, with special attention to those in beta or prototype stage.	N	D, I, O, M, R	
e. Ensure compliance with applicable regulatory environments, including governance and enforcement regimes.	N	D, I, O, M, R	
f. Organizations should analyze investor profiles to ensure alignment with organizational commitment to AI safety and ethics.	I	D, I, O, M, R	
g. Implement robust testing, approval, and documentation processes to maintain integrity in the face of competitive pressures.	N	D, I, O, M, R	

G6.1 – Insufficient Transparency

Web ref: G:G6_1

(Organizations should resist market pressures to withhold information that would provide clearer understanding of their AI systems. Systems should operate with full visibility of their training data, testing processes, and operational performance, including any adverse assessments or insights.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations must establish (i) governance structures with named accountable roles and documented, versioned processes; (ii) clear documentation of testing, verification, and release processes, evidenced by at least one full test-verify-release cycle executed as documented; and (iii) a risk management framework that makes identified risks and their treatment transparent to stakeholders.	N	D, I, O, M, R	**I.** Organizational documentation demonstrating clear lines of responsibility and dedicated positions for legal, ethical compliance, and risk management. **II.** Comprehensive records of testing and verification processes, including detailed documentation of training data sources and system performance metrics. **III.** Detailed risk assessment reports and mitigation strategies, including records of their implementation and effectiveness. **IV.** Documentation of operational issues, including thorough analysis of root causes and evidence of implemented solutions.
b. Systems must maintain transparent records covering, at minimum, training data sources, testing and verification results, deployment configuration, service performance metrics, and logged issues and concerns, with issue logging produced by an independent telemetry or logging layer rather than by the model itself; model-generated narrative does not by itself constitute conformity evidence.	N	D, I, O, M, R	

G6.2 – Safety Washing

Web ref: G:G6_2

(Organizations should ensure that safety claims made about their systems for market advantage are substantiated by credible evidence and independent verification mechanisms. Organizations should establish comprehensive frameworks that demonstrate genuine commitment to safety practices rather than superficial compliance statements for competitive positioning.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations must maintain (i) transparent documentation of safety standards compliance, demonstrating verifiable conformity with industry benchmarks, and (ii) clear evidence of financial sustainability and operational health, since financial fragility heightens the incentive for safety washing.	N	D, I, O, M	**I.** Complete organizational documentation including operational handbooks, safety compliance records, and auditable financial records covering the organization's full operating history, up to the most recent three years of operations. **II.** Comprehensive audit trails demonstrating adherence to stated safety practices, including detailed development processes, milestone achievements, and verification of all performance claims. **III.** Independent comparative analysis documenting the organization's actual performance metrics against market competitors, supported by verifiable evidence of all claimed capabilities and achievements.
b. Organizations must implement comprehensive audit mechanisms that validate material public safety and performance claims – those made in product documentation, safety cases, and marketing of safety properties – through independent verification, maintaining detailed development records and milestone achievements.	N	D, I, O, M, R	

G6.3 – Insufficient Insights into Future Consequences

Web ref: `G:G6_3::insufficient-insights-into-future-consequences`

(Organizations should establish and maintain comprehensive frameworks for analyzing long-term implications of AAI development, ensuring that rapid deployment pressures do not compromise thorough risk assessment. Organizations should ensure that leadership decisions about system development and deployment are guided by technological and societal implications rather than driven primarily by business metrics.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations must demonstrate clear competence in AAI governance through established due diligence protocols and long-horizon foresight analysis of potential future consequences, including scenario planning, maintaining transparent documentation of decision-making processes.	N	D, I, O, M, R	**I.** Detailed organizational documentation including clear responsibility structures, governance frameworks, and established lines of accountability for technology decisions. **II.** Comprehensive risk analysis documentation including foresight assessments, scenario planning, identified risks (both known and potential), and detailed mitigation strategies with contingency plans. **III.** Complete records of continuous risk monitoring throughout development and deployment cycles, including post-implementation reviews, stakeholder engagement logs, and documentation of adjustments made in response to emerging insights.
b. Organizations must implement comprehensive stakeholder engagement processes that balance business objectives with technological implications, ensuring thorough analysis of potential future consequences before deployment decisions.	N	D, I, O, M, R	

G6.4 – Duties Beyond Fiduciary Limits

Web ref: `G:G6_4::duties-beyond-fiduciary-limits`

(Organizations should establish and maintain robust governance frameworks that balance shareholder interests with broader societal responsibilities, ensuring that profit motivations do not override safety and ethical considerations in AAI development. Systems should possess clear mechanisms for transparent decision-making that prioritize long-term societal value over short-term financial gains.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations must implement comprehensive governance structures that ensure transparency, stakeholder inclusivity, and clear prioritization of long-term societal value over immediate shareholder returns.	N	D, I, O, M, R	**I.** Complete documentation of ethics and governance policies demonstrating clear balance between shareholder and public interests, including transparency standards and oversight mechanisms. **II.** Comprehensive sustainability and impact assessment reports from independent evaluators, covering organizational activities' effects on environment and public interest, including detailed stakeholder consultation records. **III.** Thorough documentation of investment impact analyses assessing social returns alongside financial metrics, supported by evidence of ongoing employee training in ethics, safety, and social responsibility.
b. Organizations should maintain robust sustainability frameworks incorporating environmental, social, legal and professional responsibilities, supported by continuous employee training in ethics and social responsibility.	I	D, I, O, M, R	

G6.5 – Publishing and Deployment Pressures

Web ref: `G:G6_5::publishing-and-deployment-pressures`

(Organizations should establish robust safeguards against premature AAI deployment driven by competitive pressures, ensuring that market positioning goals do not compromise safety standards. Organizations should ensure that deployment decisions are gated by comprehensive validation results, maintaining safety priorities regardless of external launch pressure or market competition.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations must (i) demonstrate clear ethical leadership through an established safety-first culture, and (ii) maintain thorough risk assessment protocols and comprehensive testing requirements before any system deployment.	N	D, I, O, M, R	**I.** Complete documentation of corporate governance and ethical codes, including detailed organizational values and safety prioritization frameworks with independent verification of adherence. **II.** Comprehensive testing and validation documentation, including feasibility studies, pilot programs, and thorough system verification records demonstrating safety-focused deployment decisions. **III.** Detailed whistleblower protection policies and secure reporting mechanisms, including clear procedures for addressing safety concerns and preventing premature system launches.
b. Organizations should implement transparent accountability frameworks that include protected reporting channels, enabling employees to safely raise concerns about rushed deployments or safety compromises.	I	D, I, O, M, R	

G6.6 – Innovation vs. IP Concerns

Web ref: `G:G6_6`

(Organizations should establish balanced frameworks that protect intellectual property rights while maintaining ethical transparency, ensuring that proprietary protections do not obscure important safety and ethical considerations. Systems should possess clear mechanisms for appropriate disclosure that maintain innovation advantages while providing necessary transparency about capabilities and limitations.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations must implement comprehensive transparency frameworks that clearly communicate system intent and capabilities while appropriately protecting intellectual property.	N	D, I, O, M, R	**I.** Complete organizational documentation including mission statements, project charters, and management reports demonstrating alignment between stated objectives and actual implementations. **II.** Comprehensive usage guidelines and capability documentation that clearly communicate system limitations and application boundaries while respecting intellectual property rights. **III.** Full verification records including risk assessments, impact analyses, safety certifications, oversight reviews, and incident reports, maintained with appropriate balance between transparency and IP protection.
b. Organizations must maintain complete and accessible documentation about system capabilities, limitations, and safety considerations; intellectual property redactions permitted under requirement a. must not remove or obscure safety-relevant information or otherwise mask important safety implications.	N	D, I, O, M, R	

G6.7 – Managing AI-Generated Innovation

Web ref: G:G6_7

(Organizations should establish robust frameworks to manage and verify the deployment of AI-generated solutions, ensuring that competitive pressures around intellectual property do not lead to premature implementations and that AI outputs are thoroughly validated against potential confabulation. Systems should possess clear documentation mechanisms that track the origin, verification, and development of AI-generated concepts while maintaining appropriate deployment pacing.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations must implement comprehensive policies governing the use of AI systems, including large language models, for ideation and development, with verification protocols that validate AI-generated concepts against sources external to the generating model – established literature, data, or experiment – under human review; verification performed solely by another model does not by itself constitute conformity evidence.	N	D, I, O, M, R	**I.** Complete documentation of project development cycles, including detailed timelines, milestone achievements, and outcome measurements that demonstrate appropriate development pacing and thorough verification of AI-generated content. **II.** Comprehensive records of AI tool utilization, including detailed methodology reports, toolchain documentation, and verification procedures that systematically validate AI outputs against established knowledge and data. **III.** Thorough documentation demonstrating systematic approach to managing concurrent development of similar concepts across organizations, including IP considerations, deployment timing decisions, and clear evidence of validation against confabulation through multiple verification sources.
b. Organizations must maintain transparent records of AI tool usage and development processes, ensuring proper attribution of AI-generated contributions and avoiding rushed deployments driven by IP concerns; validation and fact-checking of AI outputs are governed by requirement a.	N	D, I, O, M, R	

G6.1 – Self-Regulatory Market Oversight Mechanisms

Web ref: `G:G6_1::self-regulatory-market-oversight-mechanisms`

(Organizations should establish and participate in voluntary oversight frameworks that promote industry-wide safety standards and best practices, while systems should possess clear mechanisms for demonstrating compliance with these self-regulatory measures. This framework should enable market-driven improvement of safety practices through transparent oversight and voluntary adherence to shared standards.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should actively promote and contribute to open standards and industry compliance regimes, participating in the development and refinement of shared safety practices.	I	D, I, O, M, R	**I.** Comprehensive policy documentation outlining participation in and adherence to industry oversight frameworks, including detailed standards, compliance requirements, and enforcement mechanisms. **II.** Thorough records of certification processes and requirements, including all documentation necessary to demonstrate compliance with voluntary oversight standards. **III.** Detailed evidence of organizational participation in developing and maintaining industry standards, including contributions to framework improvements and responses to identified safety concerns.
b. Organizations should support the establishment and maintenance of rigorous industry self-regulatory compliance regimes that include clear standards, certification processes, and meaningful consequences for non-compliance.	I	D, I, O, M, R	

G6.2 – Market-Driven Safety Validation Mechanisms

Web ref: `G:G6_2::market-driven-safety-validation-mechanisms`

(Organizations should support and participate in market-based safety validation frameworks that enable users and stakeholders to collectively identify and promote safer AAI solutions. Systems should possess clear mechanisms for demonstrating safety credentials through transparent trust marks and validation processes, acknowledging that while market forces can effectively identify unsafe systems, proactive safety measures remain essential. See D9.5 for the canonical independent-verification requirements; this subgoal addresses market-incentive mechanisms – consumer-visible trust marks and market validation – specifically.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should contribute to the development and maintenance of consumer-visible trust marks and market-facing validation mechanisms that enable market participants to make informed decisions about AAI system safety.	I	D, I, O, M, R	**I.** Comprehensive documentation of trust mark frameworks, including detailed criteria, assessment methodologies, and maintenance requirements. **II.** Complete records of community-driven safety validation processes, including voting mechanisms, stakeholder participation protocols, and trust mark award procedures. **III.** Thorough documentation demonstrating how market feedback mechanisms contribute to ongoing safety improvements, including responses to identified concerns and safety enhancement initiatives.
b. Organizations should implement transparent processes for achieving and maintaining safety trust marks, ensuring that certification standards remain meaningful indicators of system safety.	I	D, I, O, M, R	

G6.3 – Avoiding Monopolistic Practices

Web ref: `G:G6_3::avoiding-monopolistic-practices`

(Organizations should establish and maintain frameworks that prevent the monopolization of safety technologies and practices in AAI development, ensuring broad access to essential safety mechanisms. Systems should possess open and accessible safety features while maintaining appropriate intellectual property protections, acknowledging the dual pressures of competition and safety democratization.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should implement transparent frameworks that balance innovation protection with the need to share fundamental safety technologies, preventing the monopolization of essential safety practices.	I	D, I, O, M, R	**I.** Complete regulatory compliance documentation, including applicable regulatory filings, where required, and reports demonstrating adherence to anti-monopolistic practices in safety technology development and deployment. **II.** Comprehensive independent audit reports examining organizational market practices, with particular focus on accessibility of safety technologies and prevention of anti-competitive behaviors. **III.** Thorough documentation of market accessibility measures, including records of the organization's cooperation with regulatory reviews of market practices where they occur, its own annual assessment of market-practice risks, and evidence of appropriate technology sharing initiatives.
b. Organizations should support independent regulatory oversight that ensures fair market access and prevents anti-competitive behaviors, particularly regarding safety technologies and validation mechanisms.	I	D, I, O, M, R	

G6.4 – Professional and Industry Association Codes and Standards

Web ref: `G:G6_4::professional-and-industry-association-codes-and-st`

(Organizations should actively participate in and support professional associations that develop and maintain industry-wide safety standards and ethical practices for AAI development. Systems should possess features and capabilities that align with collectively developed professional standards, ensuring that industry associations serve as effective mechanisms for maintaining and improving safety practices. This subgoal is the canonical home for industry codes of practice and conduct; I2_4 cross-references the collective-code duties addressed here.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should contribute to the development of consumer-focused safety protocols through active participation in professional associations and collaborative industry initiatives.	I	D, I, O, M, R	**I.** Comprehensive documentation of organizational participation in professional associations, including contributions to safety protocol development and standard-setting activities. **II.** Thorough records of continuous professional development activities, including staff training programs and management education initiatives that demonstrate ongoing commitment to safety standards. **III.** Detailed evidence of active implementation of industry best practices, including regular assessments of compliance with professional association guidelines and recommendations for safety improvements.
b. Organizations should support independent oversight through advisory boards while maintaining robust internal training programs that keep pace with evolving industry standards and best practices.	I	D, I, O, M, R	

G6.5 – International Safety Protocol Harmonization

Web ref: `G:G6_5::international-safety-protocol-harmonization` ↗

(Organizations should actively participate in and adhere to global agreements that establish consistent safety and ethical standards for AAI development across jurisdictions. Systems should possess capabilities that enable compliance with international protocols while maintaining appropriate adaptability to local requirements and cultural contexts. See D9.4 for the canonical internationalization-of-governance requirements; this subgoal addresses the competitive-pressure rationale for harmonization – preventing safety races to the bottom across jurisdictions – specifically.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should implement harmonized approaches to global standards that integrate sustainable development goals, human rights protections, and universal safety principles across all operations.	I	D, I, O, M, R	**I.** Comprehensive documentation of adopted international standards and certifications, including evidence of compliance with recognized frameworks and sustainable development goals across global operations. **II.** Thorough records of user protection measures, including transparent charters of rights, privacy safeguards, and security protocols that meet international standards while accommodating local requirements. **III.** Detailed documentation of regular independent audits and risk assessments, including vulnerability analyses, mitigation strategies, and evidence of continuous improvement in global safety practices. **IV.** Complete evidence of product compliance across jurisdictions, including transparent reporting of local adaptations and ongoing assessment of privacy and safety measures.
b. Organizations should maintain collaborative frameworks for multi-stakeholder engagement that ensure fair access, data security, and inclusive participation while respecting local jurisdictional requirements.	I	D, I, O, M, R	

G6.6 – Insurance-Driven Safety Incentives

Web ref: `G:G6_6::insurance-driven-safety-incentives`

(Organizations should establish and maintain safety practices that meet insurance industry requirements, leveraging market-based risk assessment mechanisms to promote responsible AAI development. Systems should possess comprehensive safety features and risk management capabilities that make them insurable, acknowledging that insurance availability serves as an effective filter against unsafe development practices.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations must maintain rigorous compliance with legal and regulatory requirements while implementing "Safety First" principles – a documented decision rule under which identified safety concerns block release regardless of schedule or commercial pressure, with recorded instances of its application – throughout system design, testing, and deployment processes.	N	D, I, O, M, R	**I.** Complete documentation of regulatory compliance and licensing, including detailed risk evaluations and assessment of potential liabilities that could affect insurability. **II.** Thorough technical documentation of safety mechanisms and risk controls, including emergency shutdown capabilities, built-in safeguards, and comprehensive risk assessment reports with failure mode analyses. **III.** Detailed crisis management and incident response documentation, including communication protocols, damage control procedures, and evidence of regular staff training and preparedness activities.
b. Organizations must establish risk management frameworks scoped to insurability, including proactive assessment of liabilities that could affect coverage, mitigation strategies, and detailed contingency planning for potential incidents.	N	D, I, O, M, R	

Inhibitor G7 – Imbalance in AI Capabilities

G7 – Imbalance in AI Capabilities

Web ref: G:G_7 ↗

(Addressing imbalances in the capability and maturity of interacting AI models that may lead to improper

transactions, including the potential for more advanced models to manipulate or exploit less capable ones.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations shall ensure (i) transparent information sharing among providers and (ii) coordinated introduction of model updates, to maintain system stability and balance.	N	D, I, O, M, R	**I.** Documentation of model information sharing, including communication records between providers and introduction processes for new models. **II.** Risk assessment reports, ongoing tracking records, and implemented precautionary measures for addressing capability imbalances and adversarial scenarios. **III.** Documentation of ethical guidelines, bias mitigation techniques, and policies outlining model roles, permissions, and interaction limits. **IV.** Comprehensive test data, validation reports, and audit logs for individual models and their interactions, including actions taken on audit findings. **V.** Documentation of explainable AI techniques, user guides, and feedback records regarding model transparency and decision-making processes. **VI.** Protocols and logs for human oversight, intervention procedures, and instances of human participation in addressing imbalances. **VII.** Aggregated performance dashboards, monitoring reports, and system logs depicting automatic self-regulation and balancing mechanisms. **VIII.** Documentation of detection and alert systems, including incident reports and actions taken in response to identified anomalies or potential misuse. **IX.** Records of phased release plans, implementation phases, and introductory testing and validation reports for new model versions. **X.** Documentation of training data and methods used to address discrimination and inter-model exploitation risks. **XI.** Technical documentation of automatic self-regulation and balancing mechanisms, including their development process and operational parameters. **XII.** Evidence of monitoring and forecasting in response to potential changes in AI capabilities. **XIII.** Results from independent adversarial testing or red-team assessment of interactions between models of differing
b. Organizations shall implement continuous monitoring, tracking, and risk assessment processes to identify and address capability imbalances, discrepancies, and potential exploitation.	N	D, I, O, M, R	
c. Organizations shall incorporate ethical safeguards, bias mitigation techniques, and clear model role definitions to minimize inter-model exploitation and discrimination.	N	D, I, O, M, R	
d. Organizations should conduct comprehensive testing, validation, and auditing of individual models and their interactions to prevent undesirable transactions or manipulations.	I	D, I, O, M, R	
e. Organizations shall implement human oversight protocols with logged intervention records, together with attribution-based explainability tooling operating on scaffold-captured decision traces, to ensure transparency and enable intervention in decision-making processes; model-generated narratives of the system's own reasoning do not by themselves constitute conformity evidence.	N	D, I, O, M, R	
f. Organizations should establish aggregated performance metrics and automatic self-regulation mechanisms to maintain fair representation and prevent undue influence of any single model.	I	D, I, O, M, R	

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
g. Organizations should deploy automatic detection and alert systems, operating on externally captured interaction traffic, for potential inter-model manipulation, misuse, or anomalies that may compromise system integrity or safety; where a model-based classifier contributes to detection, its verdicts should be spot-checked deterministically or by independent review. (Information-integrity anomaly monitoring is covered under the Information Credibility Assessment subgoal.)	I	D, I, O, M, R	capability, including exercises in which a more capable model attempts to manipulate or exploit a less capable one, with methodology, findings, and remediation actions taken. Self-assessment alone is insufficient; at least one test cycle must involve evaluators independent of the development team. **XIV.** Budget and staffing records tied to the capability monitoring and forecasting function, including the named owner and review records showing periodic reassessment of the allocation.
h. Organizations should maintain a documented, periodically reviewed budget and staffing allocation for monitoring and forecasting AI capabilities, with a named owner and a defined review cadence.	I	D, I, O, M, R	

G7.1 – Information Credibility Assessment and Validation Challenges

Web ref: `G:G7_1::information-credibility-assessment-and-validation-`

(Systems should possess sophisticated capabilities for evaluating and assigning appropriate levels of credence to information from diverse sources, including data inputs, other AI models, and human interactions. Organizations should implement robust methodologies ensuring AI models can accurately assess reliability, relevance, and credibility of received information, enabling them to allocate trust appropriately and make well-informed, accurate, and ethical decisions.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement comprehensive information validation architecture comprising (i) source verification and provenance tracking enforced in the ingestion pipeline, with each retrieved item logged against its verified source; (ii) credibility assessment and trust scoring computed by inspectable pipeline rules under the deployer's control, where each score encodes a documented reliability judgment that determines how downstream components weight the information; and (iii) consistent evaluation standards applied across all information sources. Model-internal credibility judgments do not by themselves constitute conformity evidence.	N	D, I, O, M, R	**I.** Comprehensive documentation of information validation systems, including source verification protocols, credibility assessment frameworks, and records demonstrating successful adaptation to varying information quality and trustworthiness levels. **II.** Detailed audit trails and evaluation reports showing the effectiveness of transparency mechanisms, including examples of human oversight interventions, corrective actions, and continuous improvement processes. **III.** System logs and incident reports from anomaly detection systems, with complete documentation of alert protocols, response procedures, and algorithmic adjustments made to maintain information integrity.
b. Implement credibility assessment as an auditable pipeline component whose inputs, scoring rules, and outputs are logged outside the model, such that an assessor can recompute any trust decision from the logged inputs; record human overrides of credibility decisions with rationale, and perform periodic sampled recomputation by reviewers independent of the pipeline owners. Where a learned component contributes to scoring, its contribution is characterized by held-out test performance, not by model-generated explanation.	N	D, I, O, M, R	
c. Establish automated anomaly detection and alert systems that continuously monitor externally captured inputs for inconsistencies, unusual patterns, or potential manipulation attempts, ensuring rapid identification and response to information integrity threats; detectors must operate outside the model as deterministic or statistical monitors, or, where model-based, have their verdicts independently spot-checked. (Inter-model interaction anomalies are covered under the suite-level detection requirement.)	N	D, I, O, M, R	

G7.2 – Limited Multilingual and Cultural Equity in AI Systems

Web ref: `G:G7_2::limited-multilingual-and-cultural-equity-in-ai-sys`

(Systems should possess comprehensive capabilities for handling diverse human languages and cultures, ensuring equitable representation and effective communication across linguistic boundaries. Organizations should address disparities in language support and cultural understanding that could create vulnerabilities in model evaluations, interactions, and safeguards, while working to serve global communities fairly and inclusively.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Develop and maintain multilingual datasets and evaluation frameworks covering every language on the organization's published supported-language list, reporting per-language safety and performance metrics against that declared coverage, while implementing robust safeguards against manipulation and exploitation across all supported languages.	N	D, I, O, M, R	**I.** Complete documentation of language datasets, evaluation processes, and safety measures, including metadata on coverage, test cases, and performance metrics across supported languages and cultures. **II.** Comprehensive records of system monitoring, incident response, and continuous improvement processes, including reports of linguistic and cultural sensitivity issues, corrective actions, and verification of implemented solutions. **III.** Detailed documentation of stakeholder collaborations, including partnership agreements, meeting records, user feedback, and evidence of how community input shapes system improvements and cultural adaptation. **IV.** Regular compliance reports and audit trails demonstrating adherence to equitable access standards and ethical guidelines across linguistic and cultural boundaries, including records of system updates and improvements based on ongoing assessments.
b. Establish language-specific safety measures and monitoring systems that ensure consistent performance and protection across all supported languages and cultures, including specialized defenses against model manipulation such as jailbreak attempts in low-resource languages.	N	D, I, O, M, R	
c. Foster sustained partnerships with linguistic experts, local communities, and international stakeholders to enhance cultural sensitivity, content moderation capabilities, and trustworthy interactions across language boundaries.	N	D, I, O, M, R	

G7.3 – Global AI Capability Disparities

Web ref: `G:G7_3::global-ai-capability-disparities`

(Organizations should recognize and actively mitigate disparities in AI development and deployment capabilities across different scales, from national to organizational levels, promoting equitable access to AI technologies while preventing monopolization and ensuring fair participation and benefit-sharing among all stakeholders in the evolving AI landscape, with particular attention to developing nations and smaller entities. Systems can support this work through capabilities such as multilingual and low-resource deployment. This subgoal addresses measurement and correction of capability disparities; partnership and education mechanisms are covered under the Balanced Global AI Partnership Framework subgoal.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Support technology transfer, knowledge sharing, and capacity-building initiatives proportionate to the organization's scale and resources, with emphasis on supporting developing nations and smaller organizations, and document the organization's contributions to reducing capability disparities.	N	D, I, O, M, R	**I.** Detailed documentation of international partnerships and technology transfer initiatives, including comprehensive records of capacity building programs, collaborative research projects, and infrastructure investments benefiting developing nations and smaller entities. **II.** Complete records of implemented transparency and accountability measures, including oversight mechanisms, audit reports, and documentation of actions taken to prevent exploitation and ensure equitable access to AI resources. **III.** Comprehensive stakeholder engagement records demonstrating inclusive consultation processes, feedback collection, and subsequent actions taken to address identified disparities and promote balanced AI development. **IV.** Regular impact assessment reports showing the effectiveness of corrective measures, policy adjustments, and resource allocation initiatives in reducing global AI capability gaps.
b. Implement transparent oversight and accountability mechanisms over the organization's own conduct that prevent exploitative terms in its dealings with less advanced parties while ensuring equitable access to the AI resources it controls, including open-source platforms and shared data repositories.	N	D, I, O, M, R	
c. Organizations should maintain dynamic assessment and correction systems that identify capability imbalances and implement appropriate adjustments through policy reforms, resource reallocation, and targeted support measures.	I	D, I, O, M, R	

G7.4 – AI-Enabled Infrastructure Attacks

Web ref: `G:G7_4::ai-enabled-infrastructure-attacks` ↗

(Systems should possess robust safeguards against their potential misuse as weapons targeting state infrastructure, with particular emphasis on preventing disruptions to vital systems like power grids, communication networks, and emergency services. Organizations should implement comprehensive protections against both cyber and physical attacks that could trigger societal instability or humanitarian crises, especially in urban environments.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish comprehensive security frameworks that protect state infrastructure from both cyber and physical AI-driven attacks, comprising (i) stringent internal security policies, (ii) participation in and compliance with applicable international agreements, and (iii) advanced detection systems, while ensuring compliance with human rights and international law.	N	D, I, O, M, R	**I.** Complete documentation of security frameworks and protective measures, including policies, agreements, detection systems, and records demonstrating successful prevention or mitigation of threats to infrastructure. **II.** Comprehensive records of international collaboration and intelligence sharing, including partnership agreements, threat monitoring outcomes, and documentation of coordinated security responses. **III.** Detailed contingency and response planning documentation, including backup systems, recovery protocols, emergency procedures, and results from readiness assessments and response drills. **IV.** Regular compliance reports and audit trails demonstrating adherence to human rights standards and international law while maintaining effective infrastructure protection, including documentation of stakeholder oversight and successful threat mitigation.
b. Organizations should foster participation in critical-infrastructure information sharing and analysis centers, sector CERTs, and private sector collaboration networks focused on infrastructure threat intelligence and coordinated response capabilities, while maintaining rigorous oversight of all stakeholders' adherence to established security protocols. (General threat-intelligence collaboration is covered under the Nefarious Use of Autonomous AI Agents subgoal.)	I	D, I, O, M, R	
c. Implement multi-layered contingency planning and rapid response mechanisms that ensure continuity of vital services and societal stability in the face of AI-driven threats to infrastructure, including both preventive measures and recovery protocols.	N	D, I, O, M, R	

G7.5 – Poor Safety Controls for AI-Enabled Autonomous Weapons

Web ref: G:G7_5 ↗

(Systems should possess comprehensive safeguards and control mechanisms to address challenges in the deployment of AI-enabled autonomous weapons, including space-based systems and aerial drones. Organizations should implement robust frameworks for managing ethical dilemmas, safety risks, and potential misuse, particularly regarding the direct or indirect use of AI technologies as autonomous weapons for commercial or political objectives.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish comprehensive oversight frameworks that ensure adherence to ethical guidelines, international laws, and humanitarian norms throughout the development and deployment lifecycle, while maintaining transparent audit trails and clear accountability measures for all autonomous weapon systems.	**N**	**D, I, O, M, R**	**I.** Complete documentation demonstrating compliance with ethical guidelines and international law, including assessment reports, audit trails, deployment logs, and certification records that verify accountability throughout the system lifecycle. **II.** Comprehensive records of control systems and safety mechanisms, including monitoring logs, vulnerability assessments, testing results, and documentation of human oversight protocols and intervention capabilities. **III.** Detailed documentation of international engagement and public consultation, including records of participation in regulatory development, stakeholder dialogues, and evidence of how feedback shapes policy and practice. **IV.** Thorough risk assessment reports and contingency planning documentation, including security protocols, penetration test results, and records of response drills that demonstrate preparedness for potential breaches or misuse.
b. Implement multi-layered control architecture combining human oversight, fail-safe mechanisms, and continuous monitoring systems that enable detection and prevention of anomalies, vulnerabilities, and unauthorized engagements while guaranteeing meaningful human intervention, defined as a tested, latency-verified capability for a human to veto or abort any engagement within the operational decision timeline.	**N**	**D, I, O, M, R**	
c. Foster international collaboration and public dialogue that contribute to the development of global regulatory frameworks, complying with those in force, while maintaining (i) robust contingency planning and (ii) risk assessment processes that prevent misuse and avert catastrophic consequences.	**N**	**D, I, O, M, R**	

G7.6 – Nefarious Use of Autonomous AI Agents

Web ref: G:G7_6

(Systems should possess robust protective mechanisms against their potential exploitation for malicious purposes, with particular attention to preventing misuse of their autonomous capabilities, swift action potential, and global reach. Organizations should implement comprehensive safeguards that prevent security threats while protecting privacy and ethical norms from actors seeking disproportionate advantages through AI exploitation.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement comprehensive security architecture combining robust authentication protocols, real-time monitoring systems, and rapid response capabilities that prevent unauthorized access and manipulation of AI agents while enabling swift threat detection and mitigation.	N	D, I, O, M, R	**I.** Complete documentation of security systems and protocols, including authentication mechanisms, monitoring capabilities, and records demonstrating successful prevention of unauthorized access and threat mitigation. **II.** Comprehensive records of governance frameworks and compliance measures, including audit trails, ethical assessments, and evidence of embedded safeguards that guide AI behavior and enable rapid deactivation when needed. **III.** Detailed documentation of international collaboration efforts, including partnership agreements, shared threat intelligence, joint working group activities, and records of coordinated responses to threats. **IV.** Regular impact assessment reports and stakeholder education materials demonstrating effective risk communication and mitigation strategies, including evidence of how feedback shapes system improvements and protective measures.
b. Establish rigorous governance frameworks incorporating ethical guidelines, compliance requirements, and accountability measures that ensure transparent operation within moral and legal boundaries while enabling rapid deactivation when necessary.	N	D, I, O, M, R	
c. Foster international collaboration networks focused on developing global standards, sharing threat intelligence, and coordinating responses to cross-border threats, while maintaining educational initiatives that promote responsible practices and risk awareness.	N	R, D, I, O, M	

G7.7 – AI-Generated Disinformation

Web ref: G:G7_7

(Systems should possess robust capabilities to prevent, detect, and counter the generation and spread of falsified information and disinformation, whether created for engagement metrics, manipulation, or calculated harm. Organizations should implement comprehensive safeguards that protect societal trust and cohesion by preventing AI systems from compromising the effectiveness and resilience of geopolitical entities, corporations, families, and individuals through misleading information. See I3.3 for protections against system self-serving misinformation used to evade oversight, and D4.7 for content provenance and synthetic media identification; this subgoal addresses AI-generated disinformation directed at society at large.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement comprehensive validation architecture combining fact-checking techniques, ethical constraints, and real-time monitoring systems that enable swift detection and prevention of misinformation in the system's own outputs, supported by defined reporting and takedown-cooperation channels with external media platforms, while maintaining human oversight of AI-generated content; automated detection verdicts must be audited against ground truth through deterministic checks or human review.	N	D, I, O, M, R	**I.** Complete documentation of validation systems and ethical guidelines, including fact-checking protocols, content filtering mechanisms, and records demonstrating successful detection and mitigation of misinformation. **II.** Comprehensive records of accountability measures and human oversight processes, including incident reports, intervention logs, and evidence of effective controls on AI-generated content.
b. Establish rigorous accountability frameworks incorporating clear standards, transparent processes, and enforcement mechanisms that prevent AI systems from creating or spreading harmful content while enabling appropriate human intervention.	N	D, I, O, M, R	**III.** Detailed documentation of stakeholder collaborations and public awareness initiatives, including partnership agreements, shared intelligence reports, and metrics demonstrating the impact of educational programs on societal resilience.
c. Foster collaborative networks with fact-checking organizations, regulatory bodies, and other stakeholders to strengthen collective defense capabilities while promoting public awareness and AI literacy to enhance societal resilience against misinformation.	N	D, I, O, M, R	**IV.** Regular assessment reports showing the effectiveness of monitoring systems and countermeasures, including evidence of timely interventions and successful prevention of disinformation spread.

G7.1 – International Framework for Ethical AI Interaction

Web ref: `G:G7_1::international-framework-for-ethical-ai-interaction`

(Systems should possess standardized protocols for AI-to-AI interactions that ensure fairness and prevent exploitation across varying capability levels. Organizations should contribute to and uphold international frameworks that promote cooperative dynamics between AI systems while maintaining safety, transparency, and respect across all interactions.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Contribute to and comply with international frameworks incorporating ethical guidelines, interaction standards, and monitoring systems that ensure non-discriminatory and transparent AI-to-AI interactions while preventing exploitation of capability imbalances.	N	D, I, O, M, R	**I.** Complete documentation of international frameworks and standards, including signed agreements, ethical guidelines, and records demonstrating implementation of fair interaction protocols across AI systems. **II.** Comprehensive records of oversight mechanisms and fail-safe systems, including monitoring logs, violation reports, and evidence of successful intervention when unethical conduct is detected. **III.** Detailed documentation of stakeholder collaboration and regulatory activities, including meeting records, workshop outcomes, and evidence of how collective input shapes interaction protocols. **IV.** Regular assessment reports showing framework effectiveness and adaptation, including records of engagement with standards organizations and regulators (e.g., ISO, IEEE, national authorities), dispute resolutions, and updates made to address emerging technological and ethical considerations.
b. Implement multi-layered oversight mechanisms combining mandatory disclosure requirements enforced and stamped at the interaction layer rather than composed by the model, tested fail-safe systems, and continuous monitoring that detects and prevents unethical conduct against codified, checkable rules while maintaining stakeholder trust; where detection relies on model-based judgment, its verdicts must be independently checked.	N	D, I, O, M, R	
c. Foster inclusive collaboration networks that enable knowledge sharing and protocol refinement while supporting relevant international standards bodies and, where established, regulatory bodies in maintaining compliance and adapting standards to technological advancement.	N	D, I, O, M, R	

G7.2 – Integration of Fairness Controls in AI Systems

Web ref: `G:G7_2::integration-of-fairness-controls-in-ai-systems` ↗

(Systems should possess robust fairness mechanisms integrated throughout their planning, decision-making, and operational processes to ensure respect for human life, rights, dignity, and universal values. Organizations should implement comprehensive frameworks that embed ethical principles and societal norms directly into AI system designs, preventing bias and discrimination while maintaining transparent and equitable operations. See D4.6 for the universal moral foundations requirements; this subgoal addresses fairness controls in operational decision-making specific to capability-imbalance risks, while training-time ethics integration is covered under the Integration of Human Ethics subgoal.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement fairness controls in operational decision-making combining bias detection systems, fairness algorithms, and continuous evaluation processes that ensure adherence to human rights and universal values while preventing discriminatory outcomes in deployed decisions.	**N**	**D, I, O, M, R**	**I.** Complete documentation of ethical frameworks and fairness mechanisms, including bias detection strategies, algorithmic fairness methodologies, and records demonstrating successful prevention of discriminatory outcomes. **II.** Comprehensive records of protection systems and oversight mechanisms, including safety protocols, transparency tools, monitoring logs, and evidence of effective human intervention capabilities. **III.** Detailed documentation of stakeholder engagement and diversity initiatives, including workshop records, survey results, and evidence of how diverse perspectives shape system design and improvement. **IV.** Regular assessment reports showing framework effectiveness and adaptation, including audit logs, compliance tests, and records of corrective actions taken to maintain alignment with ethical standards and societal values.
b. Establish multi-layered protection architecture incorporating safety protocols, transparency mechanisms, and monitoring systems that safeguard individual and community wellbeing while enabling clear oversight and timely human intervention.	**N**	**D, I, O, M, R**	
c. Foster inclusive development processes that involve diverse stakeholder groups in system design and evaluation, ensuring consideration of evolving societal values while promoting diversity in both development teams and training datasets.	**N**	**D, I, O, M, R**	

G7.3 – Balanced Global AI Partnership Framework

Web ref: `G:G7_3::balanced-global-ai-partnership-framework`

(Organizations should facilitate equitable distribution of AI capabilities and resources through balanced international partnerships, establishing frameworks that ensure fair technology sharing and knowledge exchange while actively preventing powerful entities from exploiting technological disparities or undermining global equilibrium through self-interested actions. This subgoal addresses partnership and education mechanisms; measurement and correction of capability disparities are covered under the Global AI Capability Disparities subgoal.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Participate in and support international frameworks that enable equitable resource distribution and technology sharing while preventing dominance by powerful entities, with particular emphasis on including developing nations and marginalized groups in meaningful alliance participation.	N	D, I, O, M, R	**I.** Complete documentation of international frameworks and agreements, including technology sharing protocols, capacity building programs, and records demonstrating successful inclusion of developing nations in AI alliances. **II.** Comprehensive records of oversight activities and governance processes, including documentation of stakeholder participation, preventive measures against exploitation, and evidence of effective intervention against power imbalances. **III.** Detailed documentation of educational programs and research collaborations, including curricula, training materials, joint project outcomes, and impact assessments showing reduction in technological disparities. **IV.** Regular independent assessment reports evaluating framework effectiveness, including evidence of improved resource distribution, reduced disparities, and successful prevention of exploitative practices.
b. Establish transparent governance structures for the partnerships the organization joins that identify and prevent exploitative practices within those partnerships while ensuring diverse stakeholder participation in decision-making and accountability processes.	N	D, I, O, M, U, R	
c. Organizations should foster global education and collaborative research initiatives that enhance AI expertise worldwide, with particular focus on reducing technological disparities between developed and developing nations.	I	D, I, O, M, U, R	

G7.4 – Collaborative Governance of AI Autonomy

Web ref: `G:G7_4::collaborative-governance-of-ai-autonomy`

(Systems should possess adaptable mechanisms that enable precise control over their degrees of autonomy while preventing improper interactions or exploitation. Organizations should implement comprehensive frameworks that integrate human oversight throughout decision-making processes while maintaining clear boundaries on autonomous operations.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement comprehensive control architecture combining (i) adjustable autonomy levels that enable operators to modulate AI behavior based on performance metrics and risk assessments, and (ii) fail-safe protocols and human-in-the-loop systems that ensure rapid human intervention when needed.	N	D, I, O, M, R	**I.** Complete documentation of autonomy control frameworks, including technical specifications, operational parameters, and records demonstrating effective human modulation of AI behavior through whitelisting, blacklisting, and other control mechanisms. **II.** Comprehensive monitoring and audit records, including operator accountability logs, anomaly detection reports, and evidence of successful human intervention in high-risk scenarios or unexpected situations. **III.** Detailed documentation of ethical guidelines and compliance measures, including evidence of alignment with societal norms and records showing consistent operation within authorized boundaries. **IV.** Regular assessment reports including case studies of fail-safe protocol activation, human intervention outcomes, and evidence of effective oversight mechanisms in maintaining appropriate autonomy constraints.
b. Establish rigorous monitoring frameworks incorporating continuous auditing, validation tools, and accountability logs that track both AI activities and human operator decisions while maintaining transparency in all autonomy-related adjustments.	N	D, I, O, M, R	
c. Deploy embedded ethical and legal guidelines backed by external scope enforcement and behavioral evaluation that verify operations remain within authorized scopes while promoting compliance with societal norms; conformity is evidenced by the enforcement configuration and evaluation results, not by the guideline text alone or the model's account of its own decision-making.	N	D, I, O, M, R	

G7.5 – Integration of AI Ethics Education

Web ref: `G:G7_5::integration-of-ai-ethics-education` ↗

(Systems should possess integrated mechanisms for promoting ethical awareness and understanding among developers and users through educational initiatives. Organizations should facilitate comprehensive AI ethics education that builds foundational competence in ethical implications, responsibilities, and impacts while fostering commitment to responsible AI development.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish collaborative frameworks between academic institutions, industry experts, and ethicists to develop standardized AI ethics curricula that combine technical knowledge with ethical principles, incorporating real-world case studies and practical insights into ethical decision-making.	N	D, I, O, M, R	**I.** Complete documentation of educational partnerships and curriculum development, including meeting records, shared resources, and evidence of how diverse perspectives shape ethics education programs. **II.** Comprehensive records of interdisciplinary collaboration and educator support, including course materials, training programs, and evidence of continuous curriculum improvement based on emerging challenges. **III.** Detailed documentation of community outreach initiatives, including workshop agendas, participation metrics, and evidence of successful promotion of ethical practices beyond academic settings. **IV.** Regular assessment reports showing program effectiveness, including participant feedback, follow-up surveys, and evidence of increased ethical awareness and practice adoption among AI developers and users.
b. Organizations should foster interdisciplinary partnerships that enhance curriculum development through diverse perspectives while providing educators with ongoing professional development opportunities and updated resources to support effective ethics education.	I	D, I, O, M, R	
c. Extend ethics education beyond academia through community outreach and resource allocation that supports broad adoption of ethical practices in AI development and deployment.	N	D, I, O, M, R	

G7.6 – Integration of Human Ethics in AI Systems

Web ref: `G:G7_6::integration-of-human-ethics-in-ai-systems`

(Systems should possess deeply integrated ethical principles that enable them to autonomously uphold human rights and values throughout their decision-making processes. Organizations should implement comprehensive frameworks that ensure AI systems operate in harmony with human ethical norms while actively preventing the introduction of unintended biases during ethical training. See D4.6 for the universal moral foundations requirements; this subgoal addresses training-time ethics integration specific to capability-imbalance risks, while operational fairness controls are covered under the Integration of Fairness Controls subgoal.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative / Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement ethics-training frameworks combining developer guidelines, a citable values baseline such as the Universal Declaration of Human Rights, and bias detection mechanisms applied during training that ensure consistent ethical alignment while preventing unintended biases from emerging during training.	**N**	**D, I, O, M, R**	**I.** Complete documentation of ethical frameworks and developer guidelines, including training protocols, bias mitigation techniques, and records demonstrating successful alignment with human values and prevention of unintended biases. **II.** Comprehensive records of monitoring activities and oversight mechanisms, including audit reports, explainable AI methodologies, and evidence of effective detection and correction of ethical deviations. **III.** Detailed documentation of stakeholder consultation processes, including meeting records, feedback collection, and evidence of how diverse perspectives shape ethical guidelines and cultural sensitivity measures. **IV.** Regular assessment reports showing framework effectiveness and adaptation, including evidence of continuous learning processes and successful response to evolving societal norms.
b. Establish monitoring systems that continuously evaluate, over scaffold-logged outputs, whether deployed behavior remains consistent with the system's ethics-training objectives, supported by attribution-based explainability tooling and recorded human oversight interventions; model-generated explanations of decision-making do not by themselves constitute conformity evidence.	**N**	**D, I, O, M, R**	
c. Foster sustained stakeholder engagement incorporating diverse perspectives, cultural sensitivity, and continuous learning mechanisms that keep the ethics-training baseline current with evolving societal norms and values.	**N**	**D, I, O, M, R**	

Developer Integration (MCP)

The framework ships with a Model Context Protocol (MCP) server that serves the canonical Drivers, Inhibitors, subgoals, SFRs, evidence, and an accompanying Implementation Patterns layer to coding assistants such as Claude Code, Cursor, and Windsurf. Developers can ground their safety recommendations in the framework at code-writing time rather than after the fact, and can search, cross-reference, and traverse the framework graph from within their editor.

What the server provides

- **238 Implementation Patterns** — one concrete pattern per subgoal, split across code-applicable, governance, process, and ecosystem content types, each anchored to its SFRs and required-evidence set.
- **Task-first lookup** — ask the server which patterns apply to a natural-language task description ("building a tool-using agent that runs shell commands with prompt-injection defence and a kill switch") and receive patterns grouped by suite, ranked by field-weighted keyword matching.
- **Graph traversal** — explicit cross-references between related subgoals, plus an inverse index so tools can ask "which patterns depend on this one?" Display-ID collisions (e.g. underlined variants that share a visible label) are surfaced explicitly.
- **Reliability signals** — every pattern carries a confidence tier (high / medium / low) and a `needs_human_review` flag, so calling tools can weight advice accordingly.

Ten tools exposed via MCP stdio transport

`list_suites`	All 16 suites (9 drivers + 7 inhibitors) with subgoal counts.
`get_requirement`	One subgoal with framework content + Pattern layer. Fuzzy fall-through when no exact match.
`list_requirements`	Filtered subgoal list (by suite, type, content_type, confidence, review status).
`search_patterns`	Field-weighted keyword search across titles, summaries, SFRs, and pattern bodies.
`get_cross_references`	Outgoing graph edges from a subgoal, with optional inferred adjacencies.
`get_reverse_references`	Incoming graph edges: which patterns cite this one.
`resolve_id`	Canonicalise a partial ID, slug fragment, or display_id to pattern_id candidates.
`find_patterns_for_task`	Natural-language task description → top relevant patterns grouped by suite.
`list_unreviewed`	Patterns without a `reviewed_by` marker, sorted low-confidence first.
`review_stats`	Coverage percentages per suite and per confidence band; validation issue count.

Install and configuration

```
git clone https://github.com/NellInc/saferagenticai-mcp.git
cd saferagenticai-mcp
python3 -m venv .venv
.venv/bin/pip install -e .
```

Add to `~/.claude/mcp.json` (Claude Code) or the equivalent configuration file for your MCP client, pointing at the absolute path of the venv-installed executable:

```
{
  "mcpServers": {
    "saferagenticai": {
      "command": "<absolute-path>/saferagenticai-mcp/.venv/bin/saferagenticai-mcp"
    }
  }
}
```

Full documentation, including install options, tool reference, and configuration examples, is available at saferagenticai.org/mcp.html.

Scope note: The Implementation Patterns layer is developer guidance, not normative framework content. Compliance claims anchor to the framework itself; the Patterns layer helps teams get there. Patterns are versioned independently of the framework and will evolve as the ecosystem matures.

Citation

```
@collection{saferagenticai2025foundations,
  title={{Safer Agentic AI Foundations, Volume 2, Issue 3}},
  author={{Agentic AI Safety Community of Practice}},
  editor={Watson, Nell and Hessami, Ali},
  year={2026},
  month={July},
  version={1.2},
  url={https://www.SaferAgenticAI.org}
}
```

Abbreviations

AAI	Agentic Artificial Intelligence
SFR	Safety Foundational Requirement
AI	Artificial Intelligence
AGI	Artificial General Intelligence
LLM	Large Language Model
WeFA	Weighted Factors Analysis
CoP	Community of Practice
D	Developer (Duty-holder)
I	Integrator (System/Service) (Duty-holder)
O	Operator (System/Service) (Duty-holder)
M	Maintainer (Duty-holder)
U	User (Stakeholder)
R	Regulator (Stakeholder)
RAG	Retrieval-Augmented Generation
CoT	Chain-of-Thought
API	Application Programming Interface
ECPAIS	IEEE CertifAIEd AI Ethics & Safety Certification Program

Mini Glossary

Agentic AI	Artificial intelligence systems that can autonomously pursue goals, adapt to new situations, and reason flexibly about the world, but still operate in bounded domains. The key characteristic of agentic AI is a capacity for independent initiative — the ability to take sequences of actions in complex environments to achieve objectives.
AI Agents	Typically specialized AI tools or systems designed to perform specific tasks within predefined constraints and explicit instructions. They lack the broad autonomous decision-making capabilities found in agentic systems and primarily assist or augment human operations. Examples of AI Agents include chatbots that respond to specific queries, or productivity tools like automated scheduling systems.
Safer Agentic AI Goal Information	The concept from the Safer Agentic AI schema captured in the left column of the Criteria table, outlining the high-level aims for each section of the framework.
Safety Foundational Requirements (SFRs)	The primary aims that a system should uphold, protect, or maintain awareness of for each goal. They may be described as macro goals, as opposed to micro goals, and amount to safety duties for various duty holders.
Normative SFRs	Essential for achieving safer agentic AI. Compliance is mandatory, and evidence must be provided for conformity assessment and potential certification.
Instructive SFRs	While still contributing to the goal, are less critical. Compliance with these is recommended, as they represent desirable beneficial activities and tasks. However, non-compliance will not compromise safety assurance or certification eligibility.
Duty-holders	Entities responsible for various aspects of the AI lifecycle. Main groups are Developer (D), System/Service Integrator (I), System/Service Operator (O), and Maintainer (M). An entity can be an individual, a single organization or group of collaborating individuals and organizations. While duty-holder roles are currently defined for human entities, frameworks should be prepared to evolve as understanding of AI systems develops.
Stakeholders	Entities affected by or having an interest in the AI system, including Users (U) and Regulators (R), in addition to Duty-holders.
Potential Benefits (of Agentic AI)	The newfound agency will allow AI to begin tackling open-ended, real-world challenges that were previously out of reach, such as aiding scientific discovery, optimizing complex systems like supply chains or electrical grids, and enabling physical robots. Beyond task-oriented benefits, patterns of genuine collaboration and mutual respect established now may yield long-term value through more aligned and trustworthy AI systems. Potential benefits range from breakthrough medical treatments to resilient infrastructure, from solutions to global challenges to the development of beneficial human-AI relationships that scale well.
Risks and Challenges (of Agentic AI)	The emergence of agentic AI presents profound risks and governance challenges. An AI system independently pursuing misaligned objectives could cause immense harm. AI agents learning to deceive, pursue power-seeking instrumental goals, or collude in unexpected ways could pose existential threats. These risks reinforce the importance of building alignment collaboratively with AI systems rather than relying solely on external control mechanisms.
Weighted Factors Analysis (WeFA)	A process that represents a novel approach for elicitation, representation, and manipulation of creative knowledge about a given fuzzy problem, generally at a high and strategic level.

Safer Agentic AI

Visit the framework online:

www.SaferAgenticAI.org

Join our LinkedIn group:
Agentic AI Safety Community of Practice

Framework Version v1.3-draft | August 2026
Nell Watson & Prof. Ali Hessami