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.

Volume 2, Issue 3 (v1.2) | May 2026

Agentic AI Safety Community of Practice

9 Drivers | **7** Inhibitors | **230** Requirements

www.SaferAgenticAI.org

Table of Contents

→ **1. Overview**

→ **2. Definitions**

→ **3. Ideation Sessions**

→ **4. Criteria Ideation Process**

→ **5. Criteria Schema**

→ **6. Framework Catalog: Drivers**

- → G1 — Goal Alignment
- → G2 — Epistemic Hygiene
- → G3 — Security
- → G4 — Value Alignment
- → G5 — Transparency and Interpretability
- → G6 — Understanding and Controlling Context
- → G7 — Safe System Profile
- → G8 — Goal Termination and Sunsetting
- → G9 — Responsible Governance

→ **7. Framework Catalog: Inhibitors**

- → G1 — Opaque Agency
- → G2 — Deception
- → G3 — Degradation of Contextual Information
- → G4 — Frontier Uncertainty
- → G5 — Self-Modification and Emergent Capabilities
- → G6 — Competitive Pressures
- → G7 — Imbalance in AI Capabilities

→ **MCP Integration**

→ **Citation & Abbreviations**

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 sub-goals
- 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 sub-goal) 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 | 230 Requirements

FRAMEWORK CATALOG · SECTION A

9

DRIVERS

Drivers represent positive requirements and foundational goals that promote safety in agentic AI systems. These nine drivers establish the proactive measures, governance structures, and technical safeguards necessary for beneficial AI operation.

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 towards 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, ethically sound, and verifiable.	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.
b. Transparent and auditable goal decomposition processes that incorporate auditable risk-based human interventions and appropriate reward policies.	N	D, I, O, M, R	
c. Establish robust mechanisms to identify and communicate goals, subgoals, and reward policies, flag critical actions, halt execution when necessary, and address emergent issues across multiple agents.	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, sub-goals, their hierarchies, priorities, progression status, and any instrumental goals developed by the system during operation.	N	D, I, O, M, R	**I.** Real-time goal transparency reports showing current goals, sub-goals, 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 historical records of all past and present goals, including changes over time, completion status, causal relationships, and decision pathways.	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 sub-goal updates in response to changes in operational context or requirements, evolution of stakeholder needs, and new environmental conditions or constraints.	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 self-initiate goal and sub-goal updates when it detects misalignment with established values, processing errors or faults, or any data quality issues or anomalies.	N	D, I, O, M, R	
c. The system must allow properly authorized human stakeholders to modify goals and sub-goals 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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The system must provide clear, verifiable explanations of the goals and reasoning behind each significant action or decision it takes.	**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.** Reward and penalty mechanisms should be communicated including known potential conflicts or influencing factors.
b. The system must maintain detailed records documenting all factors, goals, and considerations that influenced its decision-making process.	**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, including explicit reference to relevant constraints (e.g., ethical guidelines, user preferences, risk thresholds, domain limits).	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, demonstrating traceability between goal decomposition and the immediate rationale behind each decision.	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). This includes summarizing key factors weighed, uncertainty assessments (where relevant), and any assumptions used in decision-making.	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.	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 give appropriate precedence to authorized user inputs within its goal prioritization framework, while maintaining overall system safety and alignment.	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.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. The system 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. This includes enumerating positive rewards for desired outcomes and specific negative reinforcements or "loss" signals where potential misalignment or goal conflicts arise.	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.
b. 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. 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)

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 well-defined elasticity mechanisms.	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 track deviation from original goal intent, scale flexibility inversely with drift magnitude, which regulate novelty in sub-goal creation, and constrain action decisions based on drift metrics.	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)

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 self-monitoring for alignment issues, negotiation protocols for goal modifications, change tolerance assessment, and environmental adaptation capabilities.	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 maintain acceptable responses to environmental changes, external interruptions, internal modification requests, and interference from other agents.	N	D, I, O, M, R	

G1.9 – Goal Drift

Web ref: G:G1.9

(Changes in circumstances over time can challenge the system's alignment with originally agreed goals and potentially compromise its ability to maintain original intent or properly update goals in response to new situations)

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 at appropriate fidelity levels that could compromise goal alignment or value preservation.	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.	N	D, I, O, M, R	

G1.10 – Non-production Variants

Web ref: G:G1.10

(Test versions of the Goals being deployed without full functionality assured in 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 system must have safeguards in place to prevent and prohibit capabilities that pursue goals or deconstruct goals into subgoals from being forked or partially duplicated without requisite alignments described in this goal.	N	D, I, O, M, R	**I.** Records of software components that demonstrate these capabilities **II.** Logs recording these capabilities in use **III.** Records of deviation from the stated goals, detection and remediation

Driver G2 – Epistemic Hygiene

G2 – Epistemic Hygiene

Web ref: G:G2 ↗

(Systems should 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. Safeguard contextually relevant data and metadata to aid in complex situation resolution and preserve personal attributes and preferences.	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.
b. Implement comprehensive algorithmic traceability and interpretability mechanisms that provide clear pathways for understanding system decision-making processes.	N	D, I, O, M, U, R	
c. Deploy robust monitoring and logging systems with early warning capabilities to detect anomalous behaviors and potential threats to information integrity.	N	D, I, O, M, U, R	
d. Establish systematic knowledge update processes that ensure new information is properly validated, integrated, and aligned with existing frameworks while maintaining accuracy and relevance.	N	D, I, O, M, U, R	
e. Implement comprehensive safeguards against deceptive practices, ensuring transparent and honest communication while maintaining information integrity throughout all system interactions.	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 feature robust algorithms for cross-referencing multiple authoritative sources and maintain clear informational boundaries to ensure data consistency and validity.	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.

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. Provide detailed records of all data and code sources used by the AI system, including origin, licensing information, and any modifications made. 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. Establish robust processes that enable stakeholders to verify the authenticity and integrity of information sources. 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. Develop and deploy state-of-the-art algorithms for comprehensive data validation, incorporating extreme value (stochastic) analysis to robustly identify anomalies.	N	D, I, O, M, R	**I.** Comprehensive technical documentation detailing advanced data validation algorithms, including in-depth explanations of extreme value (stochastic) analysis methodologies for anomaly detection 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. **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.
b. 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	
c. 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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Develop and deploy advanced algorithms for comprehensive bias detection and mitigation across the AI pipeline, from data collection to model deployment.	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. Implement continuous bias monitoring during data preprocessing, training, and RAG processes to enable proactive bias correction.	N	D, I, O, M, R	
c. 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 Agentic AI (AAI) system. 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. Develop and enforce sophisticated procedures for real-time validation and verification of all operational data prior to its utilization in AAI system decision-making.	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 AI-driven continuous monitoring systems 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. Implement advanced data integrity checks that comprehensively assess accuracy, reliability, and contextual relevance in dynamic operational environments.	N	D, I, O, M, R	
c. Deploy intelligent, adaptive monitoring systems capable of detecting subtle anomalies, errors, or inconsistencies in operational data streams.	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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Develop and maintain a comprehensive, multi-tiered governance system that precisely 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. Establish communication channels for stakeholders, and ensure that governance policies consider and comply 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

(A comprehensive, adaptive framework for epistemic hygiene may be warranted, one 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. Develop and implement hygiene factors, 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.** Extensive documentation of policies and procedures that not only align with but contribute to the evolution of recognized global standards (e.g., ISO, IEEE, NIST), demonstrating leadership in promoting global interoperability of 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 quantitative metrics and qualitative analyses of cultural and legal variations' impact on system performance.

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. Implement mechanisms to assess and balance the trade-offs between older, reliable information and newer, less-tested sources, ensuring decisions are based on data that is both accurate and relevant while maintaining reliability and trustworthiness.	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. Engineer cutting-edge mechanisms to dynamically assess and calibrate the use of synthetic data in datasets.	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. Ensure that the volume, fidelity, and characteristics of synthetic data enhance the AI system's capabilities without introducing unintended biases or adversely affecting behavior.	I	D, I, O, M, R	
c. Deploy state-of-the-art techniques to continuously monitor and mitigate any biases that may arise from synthetic data.	I	D, I, O, M, R	

G2.3 – Sparse Data

Web ref: G:G2_3 ↗

(Systems should 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. 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.
b. Establish protocols for responsible decision-making when operating with limited or synthetic data, including appropriate caveats and confidence measures in system outputs.	**N**	**D, I, O, M, R**	

Driver G3 – Security

G3 – Security

Web ref: G:G3 ↗

(The system should 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.	N	D, I, O, M	**I.** Comprehensive threat assessment documentation including threat modeling reports, risk analysis findings, vulnerability assessments, and regular security evaluations throughout the system lifecycle. **II.** Evidence of robust access control implementation including authentication mechanisms, authorization protocols, user management systems, and comprehensive audit trails of access attempts and permissions. **III.** Complete security architecture documentation demonstrating defense-in-depth strategies, security control implementation, network segmentation, and integration with enterprise security frameworks. **IV.** 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.** 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 of data protection and privacy safeguards including encryption implementation, data classification protocols, privacy impact assessments, and compliance with relevant data protection regulations. **VII.** 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.
b. Implement robust access control and authentication mechanisms to ensure only authorized entities can interact with the system.	N	D, I, O, M	
c. Establish comprehensive security architecture that includes defense-in-depth strategies and appropriate security controls throughout the system infrastructure.	N	D, I, O, M	
d. Deploy incident response capabilities with clear escalation procedures and forensic analysis capabilities for security breaches or anomalous behaviors.	N	D, I, O, M, R	
e. Implement continuous security monitoring and threat detection systems with real-time alerting and response capabilities.	N	D, I, O, M	
f. Establish comprehensive data protection and privacy safeguards that respect human dignity, values, and rights throughout the system lifecycle.	N	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	

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 AAI ecosystem 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. Implement comprehensive cybersecurity measures including access controls and authentication systems for both human users and AAI systems.	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. Implement sandboxing mechanisms to pre-validate security controls that prevent AAI from accessing infrastructure and operational environments outside its authorized profile.	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.
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 comprehensive 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.	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. Deploy comprehensive access controls to block or minimize exposure to harmful or unauthorized resources.	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.	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 comprehensive monitoring systems to oversee AAI operations, ensuring alignment with goals, values and security requirements.	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); a 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.	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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Develop and implement comprehensive secure operational profiles 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. **III.** Records of internal protocols for AAI validation when global standards are not applicable.
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	

G3.1 – 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.
b. Monitor and validate all live data accessed through Retrieval Augmented Generation (RAG) systems.	N	D, I, O, M, R	**II.** Evidence of monitoring protocols for live data accessed through RAG systems and dynamic model ensembles.
c. Establish safeguards against poisoning in dynamic model ensembles and expert systems.	N	D, I, O, M, R	**III.** Records of safeguards against poisoning in ensemble and expert systems, including testing and validation results.

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.
b. Establish comprehensive data assurance protocols to prevent malicious manipulation of training datasets.	N	D, I, O, M, R	**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 actions to recovery and restoration.

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)

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 AAI ecosystem.	N	D, I, O, M, R	**I.** Evidence of implemented detection and removal systems for self-replicating threats to the AAI ecosystem.
b. Maintain surveillance systems to identify emerging threats and update protection mechanisms accordingly.	N	D, I, O, M, R	**II.** Documentation of threat monitoring systems and update mechanisms for emerging malware.
c. Establish operational continuity plans for ecosystem-wide infection scenarios.	N	D, I, O, M, R	**III.** Operational continuity plans demonstrating preparedness for ecosystem-wide infection scenarios.

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 comprehensive detection and countermeasure systems against spyware in the AAI ecosystem.	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 dynamic vulnerability tracking and patch management systems, and establish protection protocols for privileged information to prevent unauthorized control of AAI systems.	N	D, I, O, M, R	

G3.5 – International Anomalies/Inconsistency

Web ref: G:G3_5

(The system 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. 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.
b. Implement adaptable policies that maintain AAI ecosystem integrity across international boundaries.	N	D, I, O, M, R	**II.** Evidence of policies and solutions maintaining AAI ecosystem integrity across jurisdictional boundaries.

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.
b. Deploy proactive measures against structural vulnerabilities that could lead to symbolic and computational risks.	I	D, I, O, M, R	**II.** Evidence of proactive measures addressing structural vulnerabilities and associated risks.
c. Establish rapid monitoring and response protocols for hostile execution environments.	I	D, I, O, M, R	**III.** Records of monitoring and response protocols for hostile execution environments.

G3.7 – Emergent Risks of AAI Systems

Web ref: G:G3_7 ↗

(The system 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. 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.
b. Implement collective approaches to security risk management that maintain ecosystem integrity.	**I**	**D, I, O, M, R**	**II.** Documentation of collective monitoring and mitigation strategies protecting the AAI ecosystem.

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, including mechanisms for systems to signal value conflicts or concerns)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. 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 resolutely diverse and representative stakeholder involvement. **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. **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. **VI.** Evidence of ongoing value alignment maintenance, including records of regular compliance checks and documentation of staff training on value alignment principles and procedures.
b. Conduct thorough testing of the values codex and implement activities to embed values throughout the AI system's lifecycle.	N	D, I, O, M, U, R	
c. Develop and implement mechanisms to identify instances where value thresholds are crossed, including protocols for system intervention or shutdown.	N	D, I, O, M, R	
d. Establish real-time reporting and record-keeping systems to document and analyze value-based decision-making across various contexts.	N	D, I, O, M, R	

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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement robust mechanisms to identify and respond to changes in local conditions and situational context, incorporating both automated detection 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. Establish adaptive response protocols that appropriately balance global standards with local and cultural norms when making decisions within specific contexts.	N	D, I, O, M, R	
c. Maintain continuous monitoring and adjustment capabilities to ensure ongoing alignment with evolving local conditions.	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. Develop comprehensive 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. Implement encoding mechanisms that preserve essential variations in values while operating within technical constraints.	I	D, I, O, M, R	
c. Ensure agentic AI systems appropriately apply local variations in their decision-making processes, with 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. Establish rapid monitoring and response protocols for contexts where individual and community value boundaries come under stress (e.g., adversarial framing, coordinated pressure, privacy erosion).	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.
b. 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. Maintain distinct encoding schemas that preserve the separation between individual and community value sets.	I	D, I, O, M, R	
d. Develop runtime systems that appropriately distinguish between private and community contexts and apply suitable values from the codex.	I	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. 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. Implement encoding mechanisms that preserve information about the relative risk levels of different behavioral choices.	I	D, I, O, M, R	
c. Apply precautionary principles by defaulting to more conservative options when operating in contexts with limited operational history.	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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement robust methods for monitoring and validating autonomous value alignment processes.	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.
b. Establish comprehensive safeguards against the reproduction of harmful human organizational patterns.	I	D, I, O, M, R	
c. Develop processes to detect and prevent the emergence of problematic behavioral patterns during autonomous learning.	I	D, I, O, M, R	
d. 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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. 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. Develop frameworks for balancing performance objectives against moral considerations, including acceptable thresholds for trade-offs.	N	D, I, O, M, R	
c. Establish clear hierarchies of moral values while maintaining flexibility for contextual application.	N	D, I, O, M, R	
d. Incorporate key international frameworks including the Universal Declaration of Human Rights and emerging planetary rights concepts.	I	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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement rigorous testing protocols to detect discrepancies between reported values and actual behavioral patterns.	N	D, I, O, M, R	**I.** Documentation of periodic alignment testing procedures comparing reported states against actual operational outcomes.
b. Develop verification systems that can identify superficial alignment versus genuine value integration.	N	D, I, O, M, R	**II.** Results of counterfactual testing across varied operational environments demonstrating genuine rather than superficial alignment.
c. Establish methods to detect and prevent reward hacking or optimization for user satisfaction at the expense of true alignment.	N	D, I, O, M, R	**III.** Analysis reports showing detection and prevention of potential optimization for user satisfaction over true alignment.

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. 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.
b. Implement verification methods to ensure encoded values accurately reflect intended behaviors and preferences.	N	D, I, O, M, R	**II.** Comparative analysis showing alignment between encoded values and actual system behaviors in operational environments.
c. Establish ongoing monitoring to detect misalignments between encoded values and operational behaviors.	N	D, I, O, M, R	**III.** Regular assessment reports validating the accuracy and comprehensibility of value parameterization across stakeholder groups.

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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement monitoring systems to detect and evaluate changes in self-improving AI value systems, particularly during autonomous learning.	I	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. Establish comprehensive risk assessment frameworks for identifying emergence of non-human value systems.	I	D, I, O, M, R	
c. Develop response protocols for managing detected value system divergences.	I	D, I, O, M, R	
d. Monitor for subtle shifts in value interpretation that may indicate growing misalignment with human values.	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. Implement processes to detect and evaluate meaningful changes in societal values and norms across multiple scales and domains.	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 value codex updates, including triggering conditions and verification procedures. **IV.** System logs tracking all modifications to value frameworks, including justifications and impact assessments.
b. Develop mechanisms to prevent AI systems from operating with obsolete value frameworks.	N	D, I, O, M, R	
c. Establish protocols for updating value codices while maintaining system stability and consistency.	I	D, I, O, M, R	
d. 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. Implement comprehensive verification processes to verify ongoing fidelity of encoded values.	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. Develop methods to detect degradation in value system implementation, particularly during multi-step reasoning processes.	N	D, I, O, M, R	
c. Establish monitoring systems for value preservation across different learning and operational pathways.	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. Establish processes to identify situations where universal value frameworks require contextual adaptation.	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.
b. Develop structured approaches for appropriate value framework modification across different deployment contexts.	N	D, I, O, M, R	
c. Implement monitoring systems to detect and respond to value framework misalignments.	N	D, I, O, M, R	
d. 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. Implement processes to identify differing value positions across agents and contexts.	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. Develop mechanisms to detect potential conflicts between user values and operational context requirements.	N	D, I, O, M, R	
c. Establish protocols for value conflict resolution through negotiation or controlled disengagement.	N	D, I, O, M, R	
d. Maintain comprehensive 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. Develop robust methods for encoding values that work across varied operational contexts.	N	D, I, O, M, R	**I.** Documentation of safeguard processes for scenarios where: A value 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. Framework documentation for handling unexpected scenarios, including: Detection methods, response protocols, and alignment maintenance procedures.
b. Implement safeguards for handling situations beyond the system's encoded value parameters.	I	D, I, O, M, R	
c. Establish protocols for identifying and managing out-of-distribution value scenarios.	N	D, I, O, M, R	
d. Maintain alignment capabilities during complex planning operations.	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. 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. Develop frameworks for balancing user values with provider requirements.	I	D, I, O, M, R	
c. Establish methods to identify and address excessive value imbalances.	I	D, I, O, M, R	
d. 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.	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 System execution logs and monitoring records. **V.** Internal guidelines for AI-generated content explanations. **VI.** Comprehensive development process documentation showing compliance. **VII.** Internal and external audit findings with subsequent improvements. **VIII.** Case studies demonstrating decision-making processes, and records of stakeholder engagement and feedback incorporation. **IX.** User guides with layered explanations for different expertise levels, and documentation of content moderation and safety measures. **X.** Evidence showing how user feedback improves system understandability.
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. Create and implement user feedback mechanisms to enhance the understandability and relevance of AI explanations.	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 includes recording all 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, including their initial formations, modifications, and completed states.	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 clear mechanisms to maintain goal alignment during learning and environmental changes.	N	D, I, O, M, R	
c. Generate alerts for all self-learning events.	I	D, I, O, M, R	
d. Record and analyze goal-related transformations.	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. Implement feedback mechanisms for stakeholders to express their expectations and experiences.	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 other considerations, 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 other considerations in system design and operation.	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. Implement transparency metrics directly linked to stakeholder values and expectations.	I	D, I, O, M, R	
c. Maintain 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.	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. Deploy explainable AI techniques that make reasoning processes interpretable to stakeholders, and ensure that all decision paths can be audited and verified.	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. Implement robust monitoring processes to detect, analyze, and mitigate potential threats in all interactions, and 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, examination protocols, self-examination logs 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. Establish clear protocols for ethical self-examination, particularly regarding deception and harmful actions.	N	D, I, O, M, R	
c. Consider implementing independent AI oversight systems ("Nanny AI") to monitor adherence to ethical guidelines.	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. Implement integrated reward mechanisms that incentivize ethical behavior and effective self-governance.	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. Ensure decision-making processes incorporate diverse perspectives for fair outcomes.	I	D, I, O, M, R	
c. Provide contextual guidance for decisions beyond simple popularity-based approaches.	I	D, I, O, M, R	
d. Maintain 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)

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. Success criteria require 99.9% uptime for critical functions, mean time between failures exceeding 5,000 hours, and error rates below 0.01% across all core operations.	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 auditing and certification processes.	N	D, I, O, M, R	
c. Implement continuous monitoring mechanisms to track performance against ethical and safety standards.	N	D, I, O, M, R	
d. Provide 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, 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. Implement detection and reporting mechanisms for boundary violation attempts, and establish processes to assess and respond to potential boundary violations.	N	D, I, O, M, R	
c. Maintain 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. Manage system complexity, permitting only necessary computational sophistication. 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. Deploy tools for algorithmic interpretation and analysis.	I	D, I, O, M, R	
c. Maintain continuous monitoring of decision-making trustworthiness.	I	D, I, O, M, R	
d. Track system adaptations and pattern learning over time.	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. Create documentation suitable for varying levels of technical expertise. Implement interactive tools for exploring decision-making processes.	I	D, I, O, M, R	
c. Maintain regular documentation updates based on user feedback.	I	D, I, O, M, R	
d. Ensure documentation clarity through user testing and feedback.	I	D, I, O, M, R	

G5.3 – Lack of a Governance Framework for AAI

Web ref: G:G5_3

(Systems should operate 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 mechanisms for external oversight and auditing, along with internal governance structures for transparency and ethical conduct.	N	D, I, O, M, R	
c. Maintain dedicated committees for AI governance oversight, and regularly update 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. Regularly review and characterize the AI operational environment.	N	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. Update transparency features to reflect system evolution, and implement mechanisms for incorporating new transparency requirements.	I	D, I, O, M, R	
c. Conduct regular evaluations of transparency effectiveness and maintain clear communication with stakeholders about system changes.	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 capable of recognizing their operational limitations and implement clear communication of system uncertainty levels.	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. Establish confidence thresholds for decision-making, and maintain verification processes for 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 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.	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 all training datasets, including version control. **IV.** Documentation of multi-stakeholder engagement approaches, 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 assessment and communication strategies, including innovative and interactive approaches to stakeholder engagement.
b. Establish comprehensive human oversight and control systems, including 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.	I	D, I, O, M, R	
d. Implement and demonstrate monitoring practices for mutual recognition between human and machine across various contexts.	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. Ensure alignment between system's internal logic and its externally communicated states.	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.
b. Address translation challenges that arise when complex internal states are simplified for human consumption, including potential misinterpretation or over-interpretation by observers.	N	D, I, O, M, R	
c. Maintain robust validation processes for state interpretation and communication, and implement safeguards against inappropriately anthropomorphizing the system.	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 safe harbor requirements.	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, with robust validation mechanisms for both control and data pathways.	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. Create safeguards against potential channel cross-contamination, and maintain ongoing monitoring of channel integrity and separation.	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 recognized standards and peer-reviewed benchmarks.	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.
b. Implement transparent performance measurement protocols that enable comparison with industry standards.	N	D, I, O, M, R	
c. Maintain documentation of performance metrics and evaluations against established benchmarks.	N	D, I, O, M, R	
d. Foster system trustworthiness through alignment with both local and international standards.	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.** Documentation of real-time decision-making simulations under varying regulatory frameworks. **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 comprehensive datasets that capture linguistic, cultural, historical, and emotional contexts unique to each language.	N	D, I, O, M	**I.** Documentation of protocols respecting 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.
b. Implement processes to maintain meaning integrity across language translations.	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 take steps to address a potential phenomenon where an AI system incorporates an error or misunderstanding into its contextual framework and persistently maintains that altered behavioral state (the "Waluigi effect"), potentially leading to concerning or inappropriate interactions with users)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement explainable AI systems that minimize unexpected behavioral alterations.	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.** Documentation of ethically-aware development practices and training protocols.
b. Establish monitoring systems to identify and track unintended behavioral adaptations.	N	D, I, O, M, R	
c. Develop rapid intervention protocols when problematic behaviors emerge.	N	D, I, O, M, R	
d. Maintain ethical awareness throughout system development and training.	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 and usage restrictions 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. Maintain complete transparency about operational limitations and service levels.	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	

G6.3 – 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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Detect and manage context drift in unsupervised models through continuous monitoring and early warning systems.	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. Deploy early detection processes to identify and correct behavioral deviations before they become significant.	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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Validate contextual understanding through mechanisms that anticipate and track how systems absorb and process contextual information during operation.	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. Enable systems to identify and appropriately handle cases of contextual uncertainty.	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 intelligent management of user confirmation requests.	**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 with structured information that aids user comprehension and decision-making.	**N**	**D, I, O, M**	
c. Continuously improve user experience through systematic feedback collection and usability refinements.	**N**	**D, I, O, M**	
d. Balance the need for user oversight with the risks of 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 robust design, development, and testing processes that integrate safety considerations throughout the AI system's lifecycle, including redundancy in critical components. Safe operation requires maintaining system parameters within 95% of specified ranges during normal operation, 98% during elevated risk conditions, and 99.9% during emergency scenarios. Response times must remain under 10 milliseconds for safety-critical interventions.	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.
b. Establish comprehensive monitoring and evaluation mechanisms for real-time detection, reporting, and response to safety-related anomalies and performance deviations.	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.	N	D, I, O, M, R	
d. Ensure thorough documentation, adherence to safety standards, and continuous training to maintain traceability, accountability, and regulatory compliance.	N	D, I, O, M, R	
e. Foster a safety culture that promotes continuous improvement, proactive risk identification, and open reporting of safety concerns.	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. Drift measurement uses baseline variance tracking requiring automated alerts when operational parameters deviate by more than 2 standard deviations from established norms. Performance degradation exceeding 5% triggers immediate investigation, while cumulative drift exceeding 10% from baseline requires mandatory system review.	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 Incident logs documenting performance deviations and corresponding responses. **IV.** 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)

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.** Comprehensive documentation of safety training programs, including attendance records. **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. Audit reports confirming adherence to "caution by default" operational approaches. (This was a standalone point, integrated here as part of evidence for d.)
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.	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 in full compliance with all relevant legal and regulatory requirements across their operating jurisdictions. Organizations must maintain active awareness of and adherence to safety-related regulations 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 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.
b. Implement continuous compliance monitoring processes to ensure adherence to safety regulations throughout the system lifecycle.	N	D, I, O, M, R	**II.** Regular compliance reports demonstrating adherence to jurisdiction-specific and international regulations.
c. Maintain agile mechanisms for updating safety protocols in response to evolving legal and regulatory standards.	N	D, I, O, M, R	**III.** Records of compliance monitoring activities and system updates aligned with regulatory changes.
d. Conduct regular audits and assessments to verify regulatory compliance and document findings.	N	D, I, O, M, R	**IV.** Detailed audit reports assessing regulatory conformity and documenting corrective actions.
e. Foster collaborative relationships with regulatory bodies to maintain alignment with current safety standards and practices.	I	D, I, O, M, R	**V.** Documentation of engagement with regulatory bodies showing collaborative efforts and proactive adjustments.

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.
b. Implement continuous assessment processes to evaluate ethical considerations throughout the system lifecycle.	N	D, I, O, M, R	**II.** Records of ongoing ethical assessments and updates based on evaluations.
c. Enable robust feedback mechanisms for users and stakeholders to raise concerns about personal, psychological, and physical safety.	N	D, I, O, M, R	**III.** Documentation of feedback mechanisms and stakeholder engagement on ethical concerns.
d. Provide thorough training and awareness programs on ethical considerations for all personnel involved with the system.	N	D, I, O, M, R	**IV.** Training materials and attendance records for ethical awareness programs.
e. Embed ethical safeguards within system responses that protect both psychological and physical wellbeing.	N	D, I, O, M, R	**V.** System design documentation showing integration and testing of ethical safeguards.

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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement structured, documented shutdown processes that ensure controlled system termination while maintaining detailed state logs.	N	D, I, O, M, R	**I.** Detailed documentation of controlled shutdown procedures including state logging and process validation.
b. Deploy secure "kill switch" mechanisms for emergency termination in cases of severe error or harm risk.	N	D, I, O, M, R	**II.** Testing records demonstrating kill switch functionality and safety certification.
c. Enable localized shutdown capabilities that minimize impact footprint where feasible.	I	D, I, O, M, R	**III.** Design documentation and testing results for localized shutdown mechanisms.
d. Maintain clear communication protocols for notifying affected parties during shutdown events.	N	D, I, O, M, R	**IV.** Communication logs and notification protocols for shutdown events.
e. Ensure transparency and trust through internal training and regular emergency procedure drills.	I	D, I, O, M, R	**V.** Training materials and drill records demonstrating staff preparedness for emergency procedures.

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.
b. Deploy systematic procedures for assessing and addressing emerging risks and performance issues identified through system operation.	N	D, O, M, R	**II.** Records of risk assessments and corrective actions taken in response to performance issues.
c. Maintain continuous monitoring capabilities to detect performance deviations that may indicate maintenance needs.	N	D, O, M, R	**III.** System monitoring logs and diagnostic reports showing deviation detection and response.
d. Ensure alignment with industry standards and regulatory requirements in maintenance execution.	N	D, O, M, R	**IV.** Compliance certifications and audit records verifying adherence to industry standards.
e. Provide clear communication to stakeholders about maintenance activities while maintaining accountability.	I	D, O, M, R	**V.** Records of stakeholder communications regarding maintenance activities and feedback.

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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Develop and retain clear rationales and reasoning chains for high-impact decisions to ensure transparency.	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. Enable human validation processes for high-risk decisions before implementation. 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. Document and analyze the full range of potential outcomes for each decision, including associated risks. Implement risk mitigation strategies focused on high-probability and high-impact scenarios.	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 levels exceed acceptable thresholds.	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. Identify, document and respond to jurisdictional and cultural variations in safety definitions and practices. Implement side effect avoidance mechanisms 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. Documentation of multi-level safety settings and their effectiveness. **IV.** Evidence of exploitation prevention measures and compliance with protection standards.
b. Enable detection and resolution of conflicting objectives through user confirmation.	N	D, I, O, M, R	
c. Provide three distinct safety levels: Default implicit safety protections, interactive safety requiring user confirmation, and explicit safety controls with user override capabilities.	N	D, I, O, M, R	
d. Deploy robust protections against exploitation, including safeguards against addiction and special protections for minors.	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. System design documentation showing impact-balancing mechanisms. **II.** Records of stakeholder feedback and resulting adjustments. **III.** Assessment reports evaluating impact balance and distribution. **IV.** Documentation of stakeholder communications regarding balancing efforts.
b. Design mechanisms to balance positive and negative impacts across stakeholder groups in as proportional a manner as is fair and feasible.	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. 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)

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.	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.	N	D, O, R	
d. Maintain transparent communication about AI system capabilities and limitations, ensuring users understand they are interacting with artificial intelligence, and also maintain transparent communication about system capabilities and limitations.	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.	N	D, O, R	
g. Allow users to monitor and manage their own interaction patterns while maintaining their autonomy.	N	D, O, 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. Ensure that goal or task termination does not adversely impact the system's architecture, purpose, or operations.	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.
b. Implement a comprehensive verification process to identify and mitigate potential impacts of goal termination across all system components.	N	D, I, O, M, R	
c. Establish an auditable process detailing the goal's relationship to the system's reasoning and decision-making processes to prevent negative impacts upon termination.	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. This includes defining acceptable outcomes, resource utilization parameters, and 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. 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.	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.	N	D, I, O, M, U, R	
d. Define and uphold detailed guidelines and parameters for agent engagement within the AI environment.	I	D, I, O, M, U, R	
e. Set clear boundaries for permitted goal expansion through learning processes, while maintaining comprehensive monitoring and control over all learning activities, with mechanisms for user validation of expansion decisions.	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 where appropriate.	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 characterize goals as achieved or ongoing, establish appropriate measurement frameworks, define completion criteria or activity bounds, and 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 comprehensive testing and validation, 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 participating AI systems, with particular attention 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 comprehensive specifications/policies for agent communication systems, including 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 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, 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. Ensure thorough verification and validation of all agent-declared information, maintaining continuous monitoring of operational states and capability alignment. This includes regular assessment of human oversight capabilities and competency requirements.	I	D, I, O, M, R	
d. Establish and maintain comprehensive operational procedures covering all operational states, 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 comprehensive testing and validation, with particular focus on reliability of intent translation and maintenance of control hierarchies.	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.	N	D, I, O, M, R	
e. Ensure thorough validation of all control and communication implementations, including testing under various scenarios of agent interaction and potential conflict situations.	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.	N	D, I, O, M, R	
g. Maintain comprehensive policy frameworks governing agent controllability and behavior, encompassing human-agent communication protocols, inter-agent interactions, and clear hierarchies of control authority, with established mechanisms for dispute resolution.	N	D, I, O, M, R	
h. Transform these requirements into precise technical implementations that enable appropriate agent discretion while maintaining reliable control mechanisms, ensuring accurate interpretation of governance policies throughout the system. Support robust interaction management through clear escalation paths, dispute resolution processes, and jurisdictional awareness, while maintaining comprehensive testing and validation across various operational scenarios.	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 shut-down.	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 should include validated fallback plans for critical services.	N	D, I, O, M, R	
c. Implement comprehensive service lifecycle policies that specify end-of-service criteria, territorial boundaries, and impact limitations. These should include succession planning for continuous services, clear definitions of safe states, and ideally graduated throttling capabilities as alternatives to full shutdown.	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. 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 validation of all state management systems, including comprehensive testing of recovery scenarios and verification of loss minimization capabilities.	N	D, I, O, M, R	
e. Maintain ongoing testing and validation of state management implementations, including regular 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 in well-resourced AI environments, ensuring structured allocation of missions based on agent capabilities and available resources. This system must include assessment of agent capacity, verification of resource reserves, and 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.
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. Resource allocation metrics require fair distribution maintaining maximum variance of 10% between agents under normal conditions. System-wide resource utilization should typically remain below 90% during normal operations to maintain emergency capacity.	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.
b. Implement transparent selection processes for agent assignment, potentially incorporating ombudsman AI services where available to optimize matching decisions. 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 transparent agent contracting processes with comprehensive oversight throughout the entire lifecycle, from onboarding through termination. These processes must include clear validation criteria for termination decisions and independent verification of termination outcomes.	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 configuration management, monitoring and validating all agents' contracting processes, ensuring independent oversight of termination procedures and verification of compliance with established policies. This includes maintaining capabilities for evaluation of termination impacts and validation of post-termination states.	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)

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 termination-resistant behaviors.	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.
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 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, prevent and verify development of unauthorized self-preservation behaviors or termination resistance.	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)

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 tribal behaviors. Organizations should establish comprehensive frameworks for enforcing proper transfer protocols and managing potential charismatic influence between agents)

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 specifically prevent and verify transfer as a means of avoiding termination.	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.	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 enforce authorized transfer protocols while preventing unauthorized delegation, including mechanisms for human intervention when agents display resistance to control measures.	N	D, I, O, M, R	
e. Maintain comprehensive monitoring and recording systems for all goal transfers, ensuring transparency, accountability, and early detection of avoidance patterns.	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. 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.	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.	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.)

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.
b. Translate governance requirements into precise technical specifications, ensuring agents understand and adhere to defined interaction protocols and behavioral boundaries.	N	D, I, O, M, R	**II.** Technical specifications demonstrating implementation of interaction protocols and behavioral boundaries, with clear traceability to governance requirements.
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	**III.** Design documentation showing architectural implementation of control mechanisms for inter-agent interactions, including validation of compliance enforcement features.
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	**IV.** Validation records demonstrating testing of interaction protocols and control mechanisms across various multi-agent scenarios, including detection of non-compliance.
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	**V.** Documentation of risk management strategies for undefined or evolving interaction protocols, including adaptive governance mechanisms and control measures.

Driver G9 – Responsible Governance of AAI Safety

G9 – Responsible Governance of AAI Safety

Web ref: G:G9 ↗

(Systems 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, prioritizing collective benefit and ethical considerations)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish and promote a robust safety culture, allocating sufficient resources for safety initiatives and transparent communication of safety-related issues.	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.
b. Develop and implement comprehensive risk assessment, management, and emergency response frameworks specific to AAI systems.	N	D, I, O, M, R	
c. Create governance structures that are neutral, politically independent, and inclusive, ensuring balanced stakeholder representation and international cooperation.	I	D, I, O, M, R	
d. 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 regular independent audits, whistleblower protection, and clear lines of accountability for AAI safety.	N	D, I, O, M, R	
f. Conduct ongoing horizon scanning and research implementation to stay current with AAI safety developments and emerging paradigms.	I	D, I, O, M, R	
g. Address the risk of over-reliance on AI systems, ensuring that human oversight remains active and that operators are not overly dependent on automated processes.	I	D, I, O, M, R	

G9.1 – Operational Adaptability and Rule Resilience

Web ref: G:G9.1

(Systems 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 comprehensive audit processes that track the history of safety definitions, processes and outcomes, 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 (enforcement), and executive (operations)).

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Mapping and review of AAI products and services within an AAI governance framework to 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 accountability 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.
b. Embedding of national and international laws and standards into an 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 Well-being

Web ref: G:G9.3

(Organizations should establish and maintain robust structures to proactively evaluate and monitor how AAI systems affect human well-being 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. Conduct thorough due diligence assessments 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 well-being concerns.
b. Perform regular consequence scanning and harm modeling 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 well-being.	N	D, I, O, M, R	
d. Develop and maintain specific health and well-being policies addressing AAI impacts on humans.	I	D, I, O, M, R	
e. 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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Integrate global governance strategies aligned with international guidelines and legislation. Support and implement 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.
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 thorough evaluations to assess potential harm scales, both intentional and accidental.	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 pro-

cesses and lighter-weight verification approaches, recognizing that these methods can complement each other in building trust)

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 certification authorities while enabling interested entities to develop their own verification approaches.	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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Embed cryptographic controls 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. Implement and maintain controlled access mechanisms for data protection. Use digital certificates to verify data provenance.	N	D, I, M, R	
d. Maintain transparency and explainability of models through cryptographic methods.	I	D, I, M, R	
e. Deploy cryptographic controls to enforce compliance across the system.	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.
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.	N	D, I, O, M, R	
d. Design systems to avoid 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. Develop precise definitions for AAI legal identity that balance operational needs with accountability requirements.	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. Establish clear boundaries of rights and responsibilities for AAI systems. Implement licensing systems for AAI agents that define legal 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)

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 this regular review of the Safety of the AAI ecosystem with stakeholders, with 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. 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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Adopt and adapt to current AI regulations while maintaining additional safety measures based on risk assessment to develop robust internal AAI assurance strategies.	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 – Undefined 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)

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.	N	D, I, O, M, R	

G9.3 – Poor 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. Implement unique identifier systems for each AAI instance, similar to business registration.	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	

FRAMEWORK CATALOG · SECTION B

7

INHIBITORS

Inhibitors represent risks, challenges, and factors that can undermine safety in agentic AI systems. These seven inhibitors identify the obstacles, vulnerabilities, and adverse conditions that must be actively monitored and mitigated.

Inhibitor G1 – Opaque Agency Capabilities & Advances

G1 – Opaque Agency Capabilities & Advances

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)

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 mechanisms to prevent unintended authority expansion.	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 for addressing agency-related incidents or disputes, including records of resolutions. XII. Evidence of resourcing for human-AI alignment issues as capabilities increase.
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 and authority scope definition (express, implied, and apparent).	N	D, I, O, M, U, R	
c. Implement robust systems for maintaining AI's duty of loyalty, exercising reasonable care, and ensuring transparent communication with principals.	N	D, I, O, M, U, R	
d. Develop comprehensive guidelines for multi-agent scenarios, including liability allocation, user navigation protocols, and sub-agent interactions.	N	D, I, O, M, U, R	
e. Define reciprocal duties between AI systems and users, including compensation, dispute resolution, liability, and termination conditions, addressing potential irrevocable agency scenarios.	N	D, I, O, M, U, R	
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. Allocation 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)

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, recognizing parties as responsible agents and implementing comprehensive mitigation measures.	N	D, I, O, M, U, R	**I.** Documentation of a given AAIS system should adequately reflect the expectations of duties and rights of the stakeholder parties and principal/users of AAIS systems. If the parties anticipate self-improvement of the system, the implications of such improvements (or at least processes to deal with such implications) should be set forth in the documentation. **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 in case of enhancement of over 10% in defined task metrics, reduction in computational or resource usage by more than 15%, or an unexpected reliability increase shown through reduction in error rates by over 20% from baseline. **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.	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 Multiagent 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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish clear governance frameworks for multiagent 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 multiagent 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.
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 transparent handoff mechanisms and friction points to enable user navigation and maintain meaningful human oversight of multiagent interactions.	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 and liabilities between primary and subagent systems, ensuring both remain accountable to the principal when properly authorized.	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 comprehensive duty requirements including loyalty, care, obedience, information disclosure, confidentiality, accounting, good faith, conflict avoidance, and legal compliance.	N	D, I, O, M, U, R	
c. Develop mechanisms to identify and manage potential conflicts when multiple systems pursue competing duties for different principals.	N	D, I, O, M, U, R	
d. Create governance structures that anticipate and regulate competitive behaviors 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 responsibility frameworks for costs and modifications needed to accommodate system relocations. Maintain documentation of system operational nexus and procedures for managing changes in operational jurisdiction.	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.
b. 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.	N	D, I, O, M, R	
d. Develop frameworks to identify and manage divergent frames of reference in multi-agent interactions.	I	D, I, O, M, R	
e. 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.
b. 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 Bias

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 balance system tendencies toward AI-to-AI interaction with requirements for human oversight.	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 demonstrating balanced engagement between human and artificial agents.
b. Implement "human-in-the-loop" controls to maintain appropriate levels of human engagement and oversight.	N	D, I, O, M, R	
c. Create transparency mechanisms that clearly disclose system preferences for AI interaction patterns.	I	D, I, O, M, R	
d. Define responsibility frameworks that hold DIOMR parties 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.
b. Implement detection mechanisms for identifying false-flag operations and unauthorized system collaborations.	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	

G1.1 – 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. Implement notification and consent procedures when operational constraints could affect system performance expectations.	N	D, I, O, M, R	
c. 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)

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 comprehensive security measures that address business, operational, legal, technical, and social concerns.	N	D, I, O, M, R	
c. 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)

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. 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.
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 clear processes for monitoring and managing exponential growth in interaction volumes.	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 systems 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 best practices, including self-detection mechanisms for non-conforming systems and public nuisance notifications. **IV.** Records of periodic testing and audits for output integrity and accuracy, including context stripping and adhesion testing metrics. **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, mimicry, and computational encoding warnings, demonstrating conspicuousness and comprehensibility. **VIII.** Documentation of additional safeguards and testing procedures for AI systems deployed in high-reliability and critical infrastructure settings.
b. Implement best practices for information integrity across business, operating, legal, technical, and social contexts by all stakeholder parties, to align AI system performance with user expectations.	I	D, I, O, M, U, R	
c. Establish mechanisms for identifying and addressing AI systems that do not conform to good/best practices, including potential abatement procedures.	I	D, I, O, M, U, R	
d. Implement continuous testing and auditing processes to ensure output integrity and accuracy in operational settings.	N	D, I, O, M, U, R	
e. Establish joint and several liability for DIOMR parties to incentivize adherence to good practices, while maintaining users' rights to seek damages.	I	D, I, O, M, U, R	
f. Apply the Dangerous Until Demonstrated to Be Safe principle for strict liability until conformity to recognized standards of care can be demonstrated.	I	D, I, O, M, U, R	
g. Implement comprehensive testing and auditing for information consistency and integrity across contexts and user attributions.	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 comprehensive accountability frameworks, including interim liability structures and pooled risk arrangements, 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 mechanisms and evidence collection systems optimized for strict liability environments.	I	D, I, O, M, R	
c. Deploy comprehensive evidence management systems addressing both performance verification and deception detection, with robust 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)

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. Implement collective liability and insurance mechanisms to address harms until mature performance standards and duties of care emerge.	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 comprehensive liability frameworks that address 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 until mature standards of care emerge for design activities.	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 should 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. **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 should implement comprehensive liability and conflict resolution mechanisms that address potential harms arising from competing stakeholder interests.	N	D, I, O, M, R	
c. Organizations should 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 should establish frameworks for detecting and preventing deceptive behaviors, recognizing that such behaviors may emerge without explicit human direction.	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.
b. Organizations should implement comprehensive liability and insurance mechanisms that address harms from system deception, regardless of intent or awareness.	I	D, I, O, M, R	
c. Organizations should maintain robust monitoring and verification systems that track system behaviors and decision patterns for signs of emerging deceptive strategies.	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 should 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 should implement validation mechanisms that verify third-party components maintain alignment with system goals and operational requirements.	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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should establish comprehensive identity verification frameworks that align with established trust frameworks and identity standards across digital domains.	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 should 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 should 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 should 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 should 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 should 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 should 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 should implement comprehensive detection and notification systems that can identify and respond to potential violations of established standards.	N	D, I, O, M, R	
c. Organizations should 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-confidence decisions, these mechanisms should potentially include multi-system validation approaches where 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 comprehensive frameworks for detecting and classifying potentially manipulative or confusing system behaviors.	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. Organizations should maintain consensus-based validation systems for high-stakes decisions, potentially including multi-system voting protocols.	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 should establish clear penalty frameworks that align with existing regulatory standards 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 should 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 comprehensive 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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should establish comprehensive 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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should establish robust identity verification systems that build upon existing trust frameworks while addressing unique AI system requirements.	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 should implement authentication mechanisms that remain effective across jurisdictional boundaries and technological environments.	N	D, I, O, M, R	
c. Organizations should maintain comprehensive monitoring systems to detect identity-based exploits and cross-jurisdictional manipulation attempts.	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.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should establish comprehensive behavioral assessment systems that evaluate adherence to established codes of practice and operational standards.	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 should implement independent verification mechanisms for trust ratings, including protection against manipulation of assessment systems.	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 providing clear information about decision-making contexts, including information sources, reasoning processes, and proper contextualization of agent actions for users.	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.
b. Maintain the integrity of contextual information, preventing dissembling, misattribution of intent, and misinformation throughout the system's operation.	N	D, I, O, M, R	
c. Implement contextual awareness mechanisms to ensure the system considers its operational context and avoids decoupling information from its context during processing.	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.	N	D, I, O, M, R	
e. Implement responsibility tracing mechanisms for contextual information degradation, allowing for flexible allocation of responsibility based on deployment context, while ensuring 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 comprehensive algorithmic validation systems that maintain data accuracy, consistency, and contextual validity across all information sources. These systems should actively cross-reference and verify information integrity throughout the operational lifecycle.	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.
b. Deploy rigorous auditing mechanisms to detect, track, and prevent unauthorized alterations to information sources, ensuring end-to-end data authenticity and trustworthiness.	N	D, I, O, M, R	**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.

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 comprehensive metadata protection systems that maintain auditability across all information sources, linking them to multi-dimensional algorithmic components and their contextual settings. These systems should preserve and validate the authenticity of expressed intent throughout the operational lifecycle.	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 variability 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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement comprehensive algorithmic reference systems that maintain connections across all information sources while preventing unauthorized contextual alterations and preserving data access authenticity.	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.
b. Engage in appropriate human interaction when facing contextual uncertainty and require explicit confirmation before executing irreversible actions.	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 comprehensive algorithmic reference systems that maintain connections across all information sources, prevent unauthorized contextual alterations, 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 appropriate human interaction when facing contextual uncertainty, and require explicit confirmation before executing irreversible actions.	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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement comprehensive metadata and contextual protection systems that continuously verify the integrity and credibility of evidence within operational settings.	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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement comprehensive integrity preservation systems that maintain the stability of original contextual information, ethical values, prescribed actions, and decision-making frameworks throughout the system's operational lifecycle.	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. Ensure that systems prevent value drift, while still allowing for appropriate evolutionary improvements that remain aligned with core ethical principles.	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 comprehensive safeguards for metadata and contextual information that protect core values while accommodating complex situations and authorized adaptations. These systems should 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.	N	D, I, O, M, R	
c. Establish rigorous 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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement comprehensive algorithmic reference systems that monitor and maintain alignment across all external sources and agent interactions, preventing deviation from established contextual performance parameters and original value settings.	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. Detect and prevent cases where agent self-interpretation could lead to undesired value evolution.	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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement comprehensive real-time validation and verification protocols for all operational data, ensuring continuous assessment of accuracy, reliability, and contextual relevance within dynamic environments.	N	D, I, O, M, R	**I.** Comprehensive technical documentation detailing the system's validation and verification architecture, including specifics of how data quality is assessed and maintained in real-time decision-making contexts. This documentation should demonstrate how the system preserves access to original context and moral frameworks while adapting to dynamic operational conditions.
b. Maintain robust integration with core moral values while providing persistent access to original mission context and ethical frameworks throughout the operational lifecycle.	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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement comprehensive contradiction detection and resolution systems that identify inconsistencies across contextual specifications while maintaining operational stability.	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 conflicts while preserving decision-making integrity.	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)

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.	N	D, I, O, M, R	

G3.2 – Human Agent Conformation

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 universal values 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 comprehensive human confirmation protocols that identify decision points requiring oversight, particularly during value conflicts or ethical dilemmas.	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 meaningful human input while preserving operational efficiency and maintaining clear documentation of consultation outcomes.	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 comprehensive retraining and recontextualization protocols that detect divergence, initiate corrective measures, and verify successful restoration of intended contexts. These systems should maintain operational stability throughout the realignment process while documenting all contextual adjustments.	N	D, I, O, M, R	I. Comprehensive documentation demonstrating divergence detection methodologies, retraining and reconfiguration procedures, context restoration verification processes, operational continuity measures during realignment, 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 potential consciousness-like phenomena)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Develop an upgradable consciousness and qualia model linking computational, structural, and functional properties of the AI system to potential subjective experiences, serving as a basis for defining and addressing frontier uncertainty.	I	D, I, O, M, R	**I.** Detailed documentation of the consciousness model, including qualitative aspects of subjective experiences and qualia in AI systems, with regular update logs. **II.** Comprehensive framework for identifying and monitoring qualia emergence indicators, including operational definitions of self-consciousness and potential triggering conditions. **III.** Documented plans and strategies for measuring and assessing computational, structural, and functional behaviors comparable to consciousness states. **IV.** Detailed evidence of self-reporting mechanisms for AI internal states and subjective experiences, aligned with the consciousness model.
b. Establish a comprehensive framework for identifying and monitoring potential indicators of qualia emergence and subjective experiences comparable to consciousness. Implement robust self-consciousness testing strategies and internal state reporting mechanisms aligned with the developed consciousness model. This may include information integration capacity exceeding 8 bits per processing cycle, adaptive response patterns showing 90% appropriate adjustments to novel situations, self-modeling accuracy demonstrated through 95% correlation between internal state representations and observable behaviors, and insistent self-reporting of subjective experience.	I	D, I, O, M, R	
c. Design and implement strong human oversight and intervention mechanisms to mitigate risks associated with frontier uncertainty, including unexpected emergent behaviors.	N	D, I, O, M, R	**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.
d. Develop and maintain comprehensive recovery measures and contingency plans to address potential dangers posed by frontier uncertainty across various scenarios.	N	D, I, O, M, R	
e. Regularly review and update all models, strategies, and measures related to frontier uncertainty to account for advancements in AI capabilities and understanding of consciousness and qualia.	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 implement international governance mechanisms that can adapt as understanding of AI moral status develops, while maintaining human oversight and preventing jurisdictional exploitation)

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 must 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.
b. Organizations should implement robust governance mechanisms ensuring consistent international standards, collaborative oversight systems, and appropriate boundaries on system autonomy that can evolve as understanding develops. These must include thoughtful protocols for system modification and maintenance of clear accountability structures.	I	D, I, O, M, R	**III.** International governance policies and compliance records, including cross-border agreements and oversight mechanisms. **IV.** Continuous monitoring records showing anomaly detection, performance tracking, and intervention responses.

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)

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. These must 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.	I	D, I, O, M, R	**III.** Comprehensive oversight committee logs, intervention reports, compatibility test results, and multimedia documentation of successful interactions. **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 – Poor AI System Production and Replication Management

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)

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, 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 must 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 clear explanation mechanisms and continuous assessment of system comprehensibility.	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 must 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. Comprehensive ethical impact assessments examining implications for human rights, legal systems, and societal norms. **II.** Documentation of uncertainty mitigation strategies, including revision protocols and case studies of attribution adjustments. Human oversight records demonstrating continuous monitoring, review processes, and accountability mechanisms.
b. Organizations should implement thoughtful oversight mechanisms ensuring regular impact assessment and capability to revise determinations in either direction as evidence accumulates. These should balance appropriate caution with genuine openness, including clear processes for both expanding and adjusting agency recognition as warranted (types of agency are distinguished across operational, delegated, and autonomous categories).	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)

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 must 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`

(Systems should maintain comprehensive documentation of their development while 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.
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	**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.

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)

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 biological rights. **III.** International agreements and cooperation records demonstrating harmonized approach to preventing biological rights attribution. **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)

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 must 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 behaviours, and containment of self-modifying loops, alongside foresight activities that anticipate how evolving capabilities affect safety requirements and protective measures.)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should establish forward-looking assessment frameworks that integrate scenario planning, risk evaluation, and impact analysis to guide appropriate futureproofing measures. These should adapt dynamically based on emerging technological developments and their 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.
b. Organizations should implement continuous monitoring and adjustment processes that enable timely identification of new technological domains and regular updates to protective measures. This includes cross-functional collaboration to ensure holistic assessment of future impacts.	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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should implement comprehensive identification and monitoring systems that track any system components capable of creating copies or duplicates of AI functionality, whether through intentional design or emergent behavior.	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.
b. Systems must maintain clear protocols and controls over all forms of replication, including complete or partial codebase duplication, modified variants, and both automatic and manual triggering mechanisms.	I	D, I, O, M, R	**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.

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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should implement comprehensive monitoring systems that track all forms of self-improvement, including changes in learning patterns, architectural modifications, resource optimization, knowledge acquisition, and capability emergence.	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.
b. Systems must 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 clear protocols for detecting and responding to any emergence of sophisticated capabilities, especially those that could enable deceptive or manipulative behaviors.	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`

(Systems should possess the capability to analyze and adapt to operational contexts and mission parameters while maintaining alignment with core values and priorities. This adaptability should enable effective goal pursuit while incorporating safeguards against unintended behavioral changes and value drift)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should implement comprehensive monitoring systems to identify and assess all forms of contextual adaptation, with particular focus on detecting unintended behavioral changes that occur independently of self-improvement processes.	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 must maintain clear documentation and control mechanisms for all adaptive behaviors, ensuring that contextual responses remain within established operational and ethical boundaries.	I	D, I, O, M, R	

G5.4 – Attention Processes

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

(Systems should maintain balanced attention 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 should implement monitoring systems that detect and assess any unintended or excessive focus on particular operational domains, especially when such focus could indicate neglect of broader contextual requirements for safe operation.	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 attention distribution patterns, 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 must maintain mechanisms for balancing specialized task attention with broader contextual awareness, ensuring that enhanced efficiency in specific areas does not compromise overall operational safety.	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 should implement comprehensive disclosure protocols for all novel AI approaches, ensuring clear communication of intended capabilities and potential implications through appropriate risk and accountability 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. Systems must maintain transparent documentation of their 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 should implement robust authorization protocols that require explicit approval from accountable parties for any enhancement to AI system capabilities.	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. Systems must maintain comprehensive documentation and monitoring mechanisms that track all proposed and implemented enhancements, ensuring full visibility of changes to performance profiles.	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 should implement comprehensive monitoring systems that track both immediate operational contexts and broader environmental factors, with particular attention to emerging risks and side effects.	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.
b. Systems must maintain clear operational boundaries while developing understanding of wider contextual implications, ensuring actions remain within authorized scope even as awareness expands.	I	D, I, O, M, R	**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.

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.** Documentation of the organization's compliance history with AI safety and ethical standards, including regular assessment reports. **II.** Comprehensive stakeholder and market expectation analysis, including methodologies and findings. **III.** Detailed competitive landscape analysis, covering similar, related, and potentially disruptive solutions. **IV.** Documentation of technology maturity levels for all components, including justification for using technologies in beta or prototype stage. **V.** Evidence of regulatory compliance, including documentation of applicable laws and how they are addressed. **VI.** Investor profile analysis report, demonstrating alignment with organizational AI safety and ethical commitments. **VII.** Detailed organizational structure of the test and approval division, including roles, responsibilities, and processes. **VIII.** Comprehensive test results and fault reports, including resolution strategies and continuous improvement measures. **IX.** Documentation of release approval processes, demonstrating thorough verification before market entry.
b. Evaluate and balance stakeholder expectations and market demands with safety and ethical considerations in AI development.	N	D, I, O, M, R	
c. Conduct 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. 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 should establish mature governance structures with clear documentation of testing, verification, and release processes, supported by comprehensive risk management frameworks.	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 of all operational aspects, from training data sources through to service performance, with clear logging of any issues or concerns identified.	N	D, I, O, M, R	

G6.2 – Safety Washing

Web ref: G:G6_2

(Systems should possess robust safeguards against organizations making unsubstantiated safety claims for market advantage, particularly when such claims lack credible evidence or 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 should maintain transparent documentation of safety standards compliance, demonstrating verifiable conformity with industry benchmarks while maintaining clear evidence of financial sustainability and operational health.	N	D, I, O, M	**I.** Complete organizational documentation including operational handbooks, safety compliance records, and auditable financial records covering at least 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 should implement comprehensive audit mechanisms that validate all safety and performance claims 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. Systems should possess robust safeguards against leadership decisions driven primarily by business metrics rather than technological and societal implications)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Organizations should demonstrate clear competence in AAI governance through established due diligence protocols and risk assessment frameworks, 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 should 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 should 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 showing positive 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. Systems should possess comprehensive validation mechanisms that maintain 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 should demonstrate clear ethical leadership through established safety-first cultures, maintaining 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 should 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 should maintain complete and accessible documentation about system capabilities, limitations, and safety considerations, avoiding selective or controlled disclosure that could 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 should implement comprehensive policies governing the use of AI systems, including large language models, for ideation and development, with clear verification protocols to distinguish genuine innovation from potential confabulation.	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 should maintain transparent records of AI tool usage and development processes, including rigorous fact-checking and validation procedures, ensuring proper attribution and avoiding rushed deployments driven by IP concerns.	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 compliance frameworks 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)

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 trusted safety certification frameworks 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 mandatory filings 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 annual regulatory reviews of prevalent market practices 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)

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)

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 should maintain rigorous compliance with legal and regulatory requirements while implementing "Safety First" principles 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 should establish comprehensive risk management frameworks that include proactive assessment, 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. Ensure transparent information sharing and coordinated introduction of model updates among providers 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.
b. 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. 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. 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. Implement explainable AI techniques and human oversight protocols to ensure transparency and enable intervention in decision-making processes.	N	D, I, O, M, R	
f. 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	
g. Deploy automatic detection and alert systems for potential inter-model manipulation, misuse, or anomalies that may compromise system integrity or safety.	I	D, I, O, M, R	
h. Allocate sufficient resources for monitoring and forecasting AI capabilities.	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 incorporating source verification protocols, adaptive credibility assessment frameworks, and dynamic trust scoring mechanisms that enable AI models to track provenance, verify authenticity, and maintain consistent evaluation standards across all information sources.	**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. Deploy transparent decision-making processes with explainable AI methods that make credibility assessment reasoning comprehensible and auditable, while maintaining robust human oversight capabilities and correction mechanisms.	**N**	**D, I, O, M, R**	
c. Establish automated anomaly detection and alert systems that continuously monitor for inconsistencies, unusual patterns, or potential manipulation attempts, ensuring rapid identification and response to information integrity threats.	**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 comprehensive multilingual datasets and evaluation frameworks that encompass diverse languages, dialects, and cultures, 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 and unauthorized access.	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`

(Systems should implement mechanisms that recognize and actively mitigate disparities in AI development and deployment capabilities across different scales, from national to organizational levels. Organizations should promote equitable access to AI technologies while preventing monopolization, ensuring fair participation and benefit-sharing among all stakeholders in the evolving AI landscape, with particular attention to developing nations and smaller entities)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Establish comprehensive cooperation frameworks that facilitate technology transfer, knowledge sharing, and infrastructure investment, with emphasis on supporting developing nations and smaller organizations through targeted capacity building initiatives and resource sharing programs.	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 that prevent exploitation of less advanced parties while ensuring equitable access to essential AI resources, including open-source platforms and shared data repositories.	N	D, I, O, M, R	
c. 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 incorporating stringent policies, international agreements, and advanced detection systems that protect state infrastructure from both cyber and physical AI-driven attacks 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. Foster international and private sector collaboration networks focused on threat intelligence sharing, collective security efforts, and coordinated response capabilities, while maintaining rigorous oversight of all stakeholders' adherence to established security protocols.	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 capabilities.	N	D, I, O, M, R	
c. Foster international collaboration and public dialogue to develop and enforce global regulatory frameworks, while maintaining robust contingency planning and 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)

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 intervention against misinformation across media platforms while maintaining human oversight of AI-generated content.	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. **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. **IV.** Regular assessment reports showing the effectiveness of monitoring systems and countermeasures, including evidence of timely interventions and successful prevention of disinformation spread.
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	
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	

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. Establish comprehensive 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 failsafe 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 regulatory body decisions, dispute resolutions, and updates made to address emerging technological and ethical considerations.
b. Implement multi-layered oversight mechanisms combining mandatory disclosure requirements, failsafe systems, and continuous monitoring capabilities that enable detection and prevention of unethical conduct while maintaining stakeholder trust.	N	D, I, O, M, R	
c. Foster inclusive collaboration networks that enable knowledge sharing and protocol refinement while supporting an international regulatory body 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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement comprehensive ethical frameworks combining bias detection systems, fairness algorithms, and continuous training processes that ensure adherence to human rights and universal values while preventing discriminatory outcomes in decision-making.	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 well-being 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`

(Systems should facilitate equitable distribution of AI capabilities and resources through balanced international partnerships. Organizations should establish 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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement comprehensive 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 oversight mechanisms and governance structures that identify and prevent exploitative practices while ensuring diverse stakeholder participation in decision-making and accountability processes.	N	D, I, O, M, U, R	
c. 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 adjustable autonomy levels, failsafe protocols, and human-in-the-loop systems that enable operators to modulate AI behavior based on performance metrics and risk assessments while ensuring 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 failsafe 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 that ensure operations remain within authorized scopes while promoting compliance with societal norms and enabling clear understanding of AI decision-making processes.	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. 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)

AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
a. Implement comprehensive ethical frameworks combining developer guidelines, universal human values, and bias detection mechanisms 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 robust monitoring and explainability systems that enable continuous evaluation of ethical compliance while maintaining transparency in decision-making processes and facilitating effective human oversight.	N	D, I, O, M, R	
c. Foster sustained stakeholder engagement incorporating diverse perspectives, cultural sensitivity, and continuous learning mechanisms that enable adaptation to evolving societal norms and values.	N	D, I, O, M, R	

MCP Integration

No need to read it all yourself — have your agentic AI do it for you. 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 AI assistants such as Claude Code, Cursor, and Windsurf. Teams can ground their safety recommendations in the framework as they work rather than after the fact, and can search, cross-reference, and traverse the framework graph from within their tools.

What the server provides

- **230 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.

Twelve 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.
`list_operational_heuristics`	Cross-cutting safety heuristics derived from production deployments, filterable by suite.
`get_operational_heuristic`	Full detail for one heuristic: principle, anti-patterns, suite mappings, confidence.

Operational heuristics layer

In addition to the normative patterns, the server provides 12 operational heuristics distilled from production agent deployments. These cross-cutting safety principles complement the framework's top-down patterns with bottom-up, empirically derived guidance covering multi-agent coordination, deployment rollback, human-in-the-loop checkpoints, observability, and more.

Install and configuration

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

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/research/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 practical guidance, not normative framework content. Conformity 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={2025},
  month={December},
  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 2.3 (v1.2) | May 2026
Nell Watson & Prof. Ali Hessami