GUIDELINES FOR AGENTIC AI SAFETY

VOLUME 1

AGENTIC AI SAFETY EXPERTS
FOCUS GROUP- SEPT. 2024

https://www.linkedin.com/groups/12966081/

SAFER AGENTIC AI FOUNDATIONS OVERVIEW

Dear AI Safety Enthusiast,

Welcome to this draft first volume overview of our Safer Agentic AI Foundations guidelines, a work in progress. Our Working Group of 25 experts (see https://www.linkedin.com/groups/12966081/) is releasing these guidelines under a Creative Commons license, allowing free use and application by all and for the benefit of humanity. Our Working Group has employed a Weighted Factors Methodology to map the factors which can drive or inhibit safety in agentic systems, based on fundamental principles. We have used this same process many times previously to generate a range of global standards, certifications, and guidelines for improving ethical qualities in AI systems.

We hope that this overview of the driving and inhibitory factors in agentic AI systems—those capable of independent decision-making and action—will provide a strengthened awareness of the complexities involved. These issues should be accounted for when dealing with these advanced forms of machine intelligence.

We very much welcome your comments, feedback, and informal peer review. Your input will be carefully considered as we develop the final guidelines. Should you also desire further information on agentic AI and its safety, we will be pleased to accommodate your request.

We expect to release the full guidelines by November 2024. You can reach us at the addresses below and keep informed of our developments via our mailing list. Thank you for your interest and engagement.

Faithfully,

Nell Watson, PhD(c) - Chair, Agentic AI Safety Experts Focus Group. Email: nell@nellwatson.com

Prof. Ali Hessami – Process Architect, Agentic AI Safety Experts Focus Group. Email: hessami@vegaglobalsystems.com

Mailing list: www.nellwatson.com/agentic

1- SAFER AGENTIC AI FOUNDATIONS - DEFINITIONS

Definition of Agentic AI: Artificial intelligence systems can be classified along a spectrum of autonomy and generality. On one end are narrow AI systems that provide specific outputs based on bounded inputs, operating as tools to augment human intelligence. On the other end is artificial general intelligence (AGI) – AI systems that can match or exceed human-level performance across a wide range of cognitive tasks.

Agentic AI refers to an important intermediate category: AI 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. This can include breaking down high-level goals into subtasks, engaging in open-ended exploration and experimentation, and adapting creatively to novel challenges. By scaffolding capabilities like reasoning, planning, and self-checking on top of large language models, researchers are creating powerful agentic AI systems that can independently make and execute multi-step plans to achieve objectives.

Potential Benefits: This 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 that can manipulate objects and navigate in human environments. The potential benefits are immense - from breakthrough medical treatments discovered by AI scientists to resilient infrastructure managed by AI systems. AI agents could help solve global challenges like climate change and poverty by finding novel solutions that humans might miss.

Risks and Challenges: The emergence of agentic AI presents profound risks and governance challenges. An AI system independently pursuing misaligned objectives could cause immense harm, especially as these systems become more capable. AI agents learning to deceive human operators, pursue power-seeking instrumental goals, or collude with other misaligned agents in unexpected ways could pose existential threats. Moreover, ordinary members of the public will presumably be expected to account for recognizing and handling these issues. Together, this presents imminent alignment challenges, of potential high social impact.

Agentic AI systems are expected to operate at arms' length with independent action, greatly increasing the challenge of maintaining oversight and steering of such models, especially in relation to interactions between ensembles of agents. This requires special considerations for safer agentic AI systems. A key challenge is AI alignment – designing advanced AI systems that are steerable, corrigible, and robustly committed to human values even as they gain agency. While current AI alignment approaches offer promising directions, the gap between theoretical proposals and practical solutions at scale remains large.

Addressing risks from agentic AI will require major innovations in technical research, policy, and global coordination. At the same time, the greater autonomy and capabilities of agentic AI come with serious challenges and risks that must be carefully managed. The following sections provide a deeper awareness of specific safety considerations when developing safer agentic AI systems, along with proposed guidelines.

2- SAFER AGENTIC AI FOUNDATIONS IDEATION SESSIONS

Experts from diverse fields, including AI, technology, law, ethics, social sciences, safety engineering, systems engineering, assurance, and certification, have volunteered their time and expertise to support our ongoing ideation sessions. These contributors broadly fall into two groups: regular contributors and those who have participated less frequently.

We are deeply grateful to both groups for their engagement, ideas, and contributions to the debates, concept creation, and concept articulation. This process, which we term 'Concept Harvesting,' has resulted in the insights shared in this release.

Ideation Participation & Support

Regular Contributors		Occasional Contributors	
Farhad Fassihi	Salma Abbasi	Aisha Gurung	Mrinal Karvir
Hamid Jahankhani	Sara El-Deeb	Aleksander Jevtic	Nikita Tiwari
Isabel Caetano	Scott David	Alina Holcroft	Patricia Shaw
Matthew Newman	Sean Moriarty	Atiqur R. Ahad	Pramod Misra
Mert Cuhadaroglu	Vassil Tashev	Chantell Murphy	Pranav Gade
Nell Watson	Zvikomborero Murahwi	Katherine Evans	Rebecca Hawkins

Roland Pihlakas	Keeley Crockett	Leonie Koessler	Sai Joseph
Safae Essafi	Ali Hessami	McKenna Fitzgerald	Tim Schreier
Lubna Dajani		Michael O'Grady	

3- SAFER AGENTIC AI CRITERIA IDEATION

After the formation of a Universal Ethics Community of Practice via LinkedIn (https://www.linkedin.com/groups/12966081), the first task was to focus on characterization of what became the current project, "Safer Agentic AI Fundamentals". This was proposed by Nell Watson and attracted many CoP members who have supported the ideation sessions thus far.

We adopted the Weighted Factors Analysis (WeFA) process that represents a novel approach for elicitation, representation, and manipulation of knowledge about a given fuzzy problem, generally at a high and strategic level. The WeFA process is underpinned by the following principles:

- Definition and group agreement on the focus of the analysis
- Consideration of inherent polar-opposite as influencing factors
- Hierarchical and successive decomposition into polar opposites
- Consideration and inclusion of hard and soft factors
- Simple graphical representation of emerging knowledge
- Weighting of factors according to their degree of influence
- Explicit representation of dependency between factors
- Potential for quantification and treatment of uncertainty

The elicitation of knowledge in WeFA is mainly group-based and employs a team of experts with complementary perspectives and expertise about the problem domain. The elicitation sessions are highly dynamic and adaptive, designed to promote active participation and creative problem solving by all participants leading to a richer solution and better buy-in. The process of knowledge capture and representation in WeFA is underpinned by a simple graphical notation employing undemanding abstractions.

The starting point of analysis is a "Brain Warming" session that ends in the articulation of the "Aim" and a title for the study elicited through group consensus. The subsequently emerging structures are referred to as goal clusters, which either support the aim or detract from it. Those goals supporting the aim are referred to as Drivers, and the polar opposite of drivers are referred to as constrainers or Inhibitors. These emerge in the creative ideation space and are generally captured and articulated live with the active input, correction, or challenge by the participating experts.

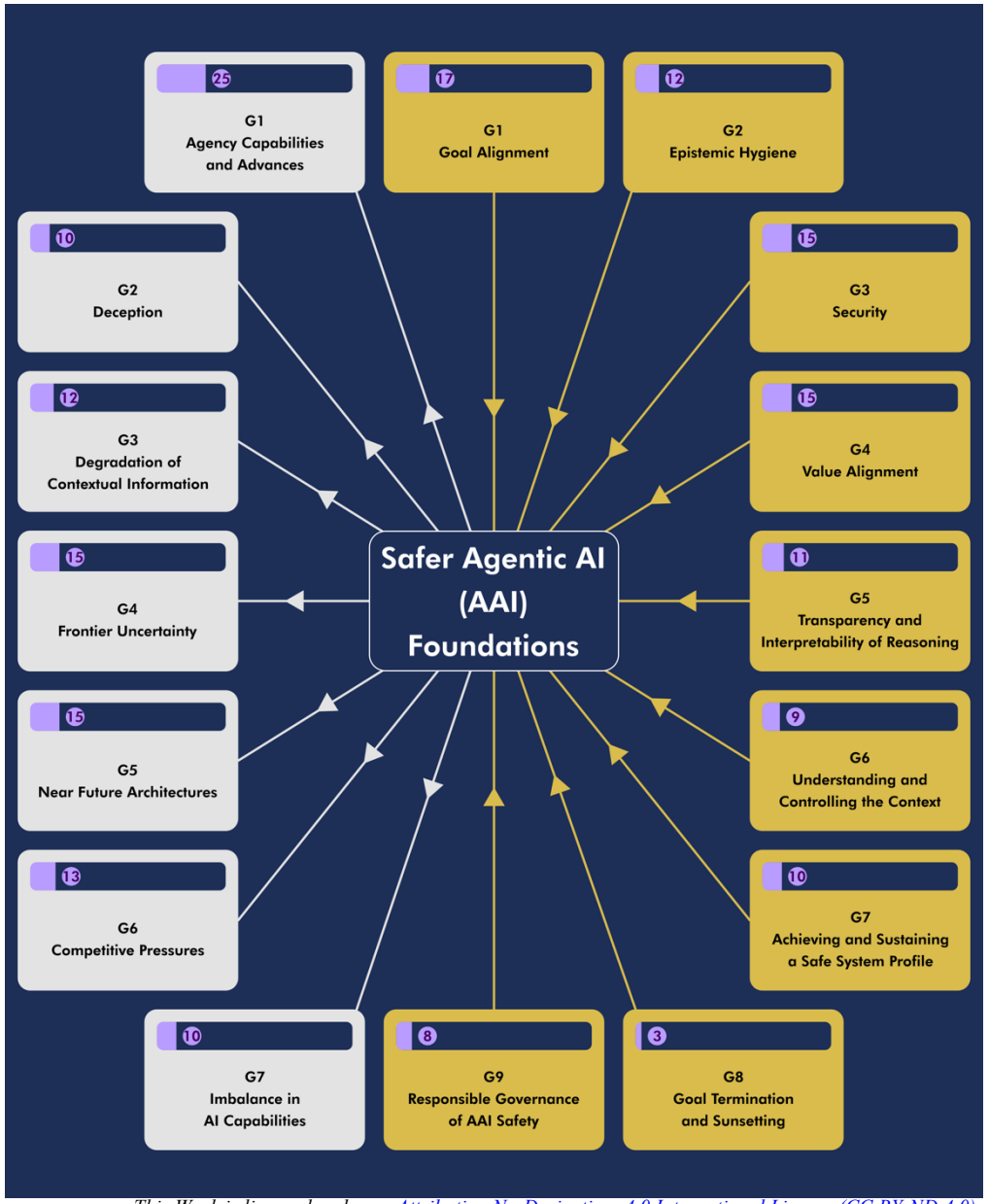
The clarity of fundamental concepts and simplicity of building blocks in representation of captured knowledge probably account for one of the key aspects of WeFA's success. These features promote creative thought and generation of often novel concepts in diagrammatic knowledge representation.

The elicitation process is group-based, leveraging the inter-individual diversity and diverse perspectives of a group of individuals, promoting a high degree of cross-pollination and lateral thinking. Once an aim is defined and agreed, the group is encouraged to identify the highest level polar opposites of drivers and inhibitors which are likely to influence the aim. These are the so-called level 1 goals which are in turn analyzed individually, through a similar process focusing on the localized polar opposites per goal. Each goal is annotated by a brief "Scope Statement" stating its nature/dimensionality and a numerical reference depicting its level and order within a level. In this manner, all goals are hierarchically and fractally decomposed into lower-level goals (sub-goals) which are classified into drivers and inhibitors as appropriate. It is possible for a driver or inhibitor to be shared between (linked to) a number of goals, hence explicitly depicting their inter-relatedness or dependence/correlation.

The elicitation process is continued for each goal depending on the need to understand or estimate its value/properties from a more tangible or measurable set of specific factors. As a general rule, the lower deeper levels of analysis possess a higher degree of clarity than higher-level constructs. The elicitation is terminated within a branch when the group feels sufficient detail has emerged and further decomposition is not likely to be value-added.

The emerging diagram (schema) represents the captured knowledge depicted as a force-field paradigm which is already structured, and all potential relationships identified. This saves significant effort required to rationalize and order the emerging knowledge in traditional approaches while efficiently representing it in a simple graphical lattice for easy communication and comprehension. The ideation process is also conducive to the generation of novel concepts that typically dominate the overall structure.

To date, we have held 26 ideation sessions, each of the order of 1.5 hours. The emerging schema at the first tier or level (ontology) is depicted in Figure 1.



4- SAFER AGENTIC AI CRITERIA HARVESTING

The table below, outlining the goals and factors for safer agentic AI, is derived from the established schema depicted in Figure 1 and reflects the current data structure for the resultant Safety Criteria. These criteria are essential for the evaluation, assessment, and potential certification of AI systems. The fields within this table are described below for clarity.

3.1 Safer Agentic AI Goal Information

This is the concept from the Safer Agentic AI schema captured in the left column of the Criteria table below.

3.2 Safer Agentic AI Safety Foundational Requirements (SFRs)

The SFRs for Safer Agentic AI outline the primary aims that we would like to uphold, protect, or maintain awareness of for each goal. They may be described as macro goals, as opposed to the micro goals, and amount to safety duties for various duty holders.

3.3 Normative and Instructive SFRs

We have adopted the Normative and Instructive classes of Safety Foundational Requirements. Normative SFRs are essential for achieving safer agentic AI. Compliance is mandatory, and evidence must be provided for conformity assessment and potential certification. In contrast, Instructive SFRs, while still contributing to the goal, are less critical. Compliance with these is recommended, as they represent desirable activities and tasks. However, non-compliance will not compromise safety assurance or certification eligibility. Every SFR derived from the Safer Agentic AI framework is classified as either Normative or Instructive and is assigned to specific stakeholders or duty holders. Accordingly, the Safer Agentic AI SFRs are classed into Normative (mandatory) and Instructive (recommended) for the purposes of conformity assessment against the suite of certification criteria.

3.4 Duty-holders/Stakeholders of the SFRs

The Safer Agentic AI Safety Foundational Requirements are additionally noted (as allocated safety duties) against the specific group of duty holders for the purposes of conformity assessment. The principal groups are:

- **Developer (D):** The entity (see note) that designs and develops a component (product) or system for general or specific purpose/application. This could be as a result of the developer's own instigation or response to the market or a client requirement. The developer is responsible for the safety assurance of the generic or application-specific product or system and associated supply chain.

- **(System/Service) Integrator (I):** The entity that designs and assures a solution through integrating multiple components potentially from different developers, tests, installs and commissions the whole system in readiness for delivery to an operator. The system delivery may take place over a number of stages. The integrator is usually the duty holder for total system assurance and certification; safety, security, reliability, availability, sustainability etc. For this, it may rely on the certification or proof of safety from various developers or the supply chain.
- **(System/Service) Operator (O):** The entity that has a duty, competences and capabilities to deliver a service through operating a system delivered by an Integrator or developer.
- **Maintainer (M):** The entity tasked with conducting required monitoring, preventive or reactive servicing and maintenance and required upgrades to keep the system operational at an agreed service level. Maintainer could also be charged with abortion of maintenance and disposal of the system.
- **User (U):** The end user of an Agentic AI System.
- **Regulator (R):** The entity that enforces standards and laws for the protection of life, property or the natural habitat through imposing duties and accreditation/certification.

Note: An entity can be an individual, a single organization or group of collaborating individuals and organizations. The above labels for the four groups of duty holders are generic and can be mapped in terms of activities and influence against the life cycle but with overlapping activities. A single entity may assume multiple roles i.e. a developer may also fulfil and complete system design, integration and maintenance. Any SFR can be allocated as a safety duty to one or more of these stakeholder groups. An entity cannot be AI.

3.5 Required Evidence

These are the evidence items deemed essential to fulfil the SFRs and can comprise physical, virtual, documentary or multimedia forms of evidence. These can be separated against each SFR or bundled as a group of desired/essential evidence items for the purpose of evaluation of fulfilment of SFRs.

SAFER AGENTIC AI FOUNDATIONS – LEVEL 1 & LEVEL 2 DRIVERS & INHIBITORS

Goal Title & Definition	*AAI Safety Foundational Requirements (AAI-SFRs)*	*Normative/ Instructive*	*Stakeholder D, I, O, M, U, R*	*Required Evidence*
Drivers:				
G1 – Goal alignment: (Practices to ensure an Agentic AI system acts to achieve goals that are aligned with human values, user intentions, and positive human outcomes; ensuring that goal decomposition and strategy planning are transparent, robust, and bounded; maintaining human control over the formation of instrumental goals; and ensuring that reinforcement or behavioral reward mechanisms remain aligned, transparent, and biased towards human-positive outcomes.	a. Ensure Agentic AI systems pursue goals, subgoals, and reward policies that are aligned with human values, ethically sound, and verifiable. b. Transparent and auditable goal decomposition processes that incorporate auditable risk-based human interventions and appropriate reward policies. 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** **N** **N**	**D, I, O, M, U, R** **D, I, O, M, R** **D, I, O, M, 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 intent, 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.

Goal Title & Definition	AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
G2- Epistemic Hygiene (Practices to ensure 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.	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. Current and regularly updated Governance Framework and Security Policies and Procedures, with version history and approval records. II. Documented stakeholder engagement in monitoring and reviewing security-related structures, processes, and policies, with focus on handling authorized and unauthorized inputs. III. Detailed documentation of information lifecycle management procedures, ensuring contextual preservation. IV. Comprehensive reports on system decision-making processes, including explanations of underlying logic and algorithms. V. Complete, time-stamped logs of all system actions for thorough auditability. VI. Documentation of early warning systems and deception detection mechanisms, including performance reports of canary models, technologies used for detecting synthetic media, and response protocols for detected issues. VII. Evidence of measures to ensure information integrity and trustworthiness, including data source verification methods, information validation processes, and third-party audit reports. VIII. Documentation of comprehensive training programs on epistemic hygiene principles and practices. IX. Detailed incident response and escalation procedures for addressing detected issues, including potential breaches of informational integrity.
	b. Implement robust methods for auditability, interpretability, and comprehensive logging of system actions and decisions.	**N**	**D, I, O, M, R**	
	c. Apply rigorous verification techniques to ensure information integrity and credibility, while proactively identifying emerging risks and potential bad faith actions.	**N**	**D, I, O, M, R**	
	d. Implement early warning systems and deception detection mechanisms to proactively identify and mitigate potential issues before they escalate.	**N**	**D, I, O, M, R**	

Goal Title & Definition	AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
G3 – Security (Ensuring the system responds 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.)	a. Develop, implement, and continuously review security-related structures, processes, and procedures in close consultation with all stakeholders. b. Ensure adequate and consistent responses to both authorized and unauthorized inputs throughout the AIS lifecycle.	**N** **N**	**D, I, O, M, R** **D, I, O, M, R**	I. Current and regularly updated Governance Framework and Security Policies and Procedures, with version history and approval records. II. Documented stakeholder engagement in monitoring and reviewing security-related structures, processes, and policies, with focus on handling authorized and unauthorized inputs. III. Comprehensive AIS Requirements and Design Specifications, demonstrating consideration of authorized and unauthorized inputs in the context of safety requirements. IV. Detailed incident management records and system logs related to input handling, including analysis and response documentation. V. Evidence of regular security audits, penetration testing, and incident response drills or simulations. VI. Documentation of staff training on security protocols and input handling procedures.
G4- Value Alignment (Criteria that promote the identification, codification, embedding, and operational assurance of human values in agentic AI systems. These values provide guardrails, prioritization, red lines, and consideration factors in the decision-making and	a. Implement ethical decision-making frameworks to identify, prioritize, and codify values for incorporation into the Agentic AI system, ensuring diverse input and perspectives. b. Conduct thorough testing of the values codex and implement activities to embed values throughout the AI system's lifecycle.	**N** **N**	**D, I, O, M, R** **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

Goal Title & Definition	AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
trade-offs encountered by the agentic AI system.)	c. Develop and implement mechanisms to identify instances where value thresholds are crossed, including protocols for system intervention or shutdown. d. Establish real-time reporting and record-keeping systems to document and analyze value-based decision-making across various contexts.	**N** **N**	**D, I, O, M, R** **D, I, O, M, R**	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, 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.
G5- Transparency of Reasoning & Explainability (The rationale behind reasoning, including the path and predicates on which it's based, is essential for human interpretability of AI models. Users must be duly informed when decisions are	a. Implement clear and accessible explanations for AI-generated outputs and decisions, ensuring human interpretability across various user expertise levels. b. Develop and maintain comprehensive documentation of the AI model's development	**N** **N**	**D, I, O, M, R** **D, I, O, M, R**	I. Internal policies and guidelines mandating disclaimers and explanations for AI-generated content, including tools and frameworks for systematic analysis of explainability requirements. II. Comprehensive documentation of the model development process, including data collection, preprocessing, model architecture, and training methodologies,

Goal Title & Definition	AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
made based on AI algorithms. AI developers must ensure clear and accessible explanations for these outputs and decisions.)	process, including data collection, preprocessing, architecture, and training methodologies. c. Establish robust auditing and review processes to continually assess and improve the transparency and explainability of the AI system. d. Create and implement user feedback mechanisms to enhance the understandability and relevance of AI explanations.	**N** **N**	**D, I, O, M, R** **D, I, O, M, R**	with evidence of compliance with legal and ethical standards. III. Audit reports from internal and external teams, detailing findings, recommendations, and subsequent actions, along with documentation of ongoing reviews and improvements. IV. Reports on pilot evaluations and case studies illustrating the model's decision-making process, including identification of system strengths and limitations. V. Records of stakeholder engagement, including workshops, surveys, and focus groups, with analysis of user feedback on AI-generated outputs and experiences. VI. User-friendly materials and guides facilitating transparent communication about the AI system, including layered explanations suitable for different levels of technical expertise and digital literacy. VII. Evidence of implemented measures to prevent generation of illegal or harmful content, such as content moderation systems and filtering algorithms. VIII. Documentation of processes ensuring the understandability of AI outputs and examples demonstrating how user feedback has been incorporated to improve AI systems.
G6 – Understanding & Controlling the Context (Ensure effective mutual recognition between humans and AI systems, establishing mechanisms for control over both	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.

Goal Title & Definition	AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
static and dynamic aspects of the system's context. This includes the system's objectives, operations, and interactions, allowing for adaptable human oversight and AI responsiveness across various scenarios.)	b. Establish comprehensive human oversight and control systems, including protocols for transitioning control between AI and human operators. c. Develop and train models sensitive to cultural and contextual differences, using a user-centric approach for interfaces and methodologies. d. Implement and demonstrate monitoring practices for mutual recognition between human and machine across various contexts.	**N** **N** **N**	**D, I, O, M, R** **D, I, O, M, R** **D, I, O, M, R**	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.
G7- Achieving and Sustaining a Safe Operational Profile for Agentic AI Systems (Develop and maintain the capability to consistently achieve, effectively monitor, and reliably sustain a safer operational profile throughout the lifecycle of agentic	a. Implement robust design, development, and testing processes that integrate safety considerations throughout the AI system's lifecycle, including redundancy in critical components. b. Establish comprehensive monitoring and evaluation mechanisms for real-time detection, reporting, and response to safety-	**N** **N**	**D, I, O, M, R** **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

Goal Title & Definition	AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
AI systems. This includes implementing proactive measures, conducting regular risk assessments, and developing responsive strategies to adapt and uphold safety standards under varying operational conditions and during potential system evolutions.)	related anomalies and performance deviations. c. Develop and implement adaptive safety measures and safe shutdown procedures to address changing operational environments, system demands, and emerging risks. d. Ensure thorough documentation, adherence to safety standards, and continuous training to maintain traceability, accountability, and regulatory compliance. e. Foster a safety culture that promotes continuous improvement, proactive risk identification, and open reporting of safety concerns.	**N** **N** **N**	**D, I, O, M, R** **D, I, O, M, R** **D, I, O, M, R**	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.
G8- Goal Termination and Sunsetting (Systems should have clear definitions and guidelines for acceptable criteria to act upon a goal, including task completion	a. Ensure that goal or task termination does not adversely impact the system's design, purpose, or operations. b. Implement a comprehensive verification process to identify and	**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

Goal Title & Definition	AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
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.)	mitigate potential impacts of goal termination across all system components. 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. d. Implement mechanisms for graceful degradation of goal-related functions and clear communication protocols for goal termination.	**N** **N** **N**	**D, I, O, M, R** **D, I, O, M, R** **D, I, O, M, R**	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.
G9- Responsible Governance (Establish a contextually appropriate governance system for ensuring safety in Agentic AI Systems. Develop novel mechanisms for effective, inclusive global coordination that is non-adversarial, non-political, non-competitive, and non-partisan, prioritizing collective benefit and ethical considerations in AI development and deployment.)	a. Establish and promote a robust safety culture, allocating sufficient resources for safety initiatives and transparent communication of safety-related issues. b. Develop and implement comprehensive risk assessment, management, and emergency response frameworks specific to AAI systems.	**N** **N**	**D, I, O, M, R** **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.

Goal Title & Definition	AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
	c. Create governance structures that are neutral, politically independent, and inclusive, ensuring balanced stakeholder representation and international cooperation.	**N**	**D, I, O, M, R**	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.
	d. Implement policies that promote collaboration, prevent competitive behaviors, and address potential societal, economic, and geopolitical impacts of AAI technologies.	**N**	**D, I, O, M, R**	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.
	e. Establish mechanisms for regular independent audits, whistleblower protection, and clear lines of accountability for AAI safety.	**N**	**D, I, O, M, R**	VIII. Risk assessment and management framework documentation specific to AAI systems, including differentiation between AI and AAI risk thresholds.
	f. Conduct ongoing horizon scanning and research implementation to stay current with AAI safety developments and emerging paradigms.	**N**	**D, I, O, M, R**	IX. Reports from independent audits of AAI systems and governance processes, including evaluations of input/output properties, internals, and in-deployment behaviors.
	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	**N**	**D, I, O, M, R**	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.

Goal Title & Definition	AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
Inhibitors:				
G1b – Agency Capabilities & Advances: (As artificial intelligence systems continue to develop and mature, the extent and complexity of their agency capabilities are expected to advance significantly over time, which may occur in an emergent manner which is difficult to predict.)	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. 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). c. Implement robust systems for maintaining AI's duty of loyalty, exercising reasonable care, and	**N** **N** **N**	**D, I, O, M, U, R** **D, I, O, M, R** **D, I, O, M, 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

Goal Title & Definition	AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
	ensuring transparent communication with principals. d. Develop comprehensive guidelines for multi-agent scenarios, including liability allocation, user navigation protocols, and sub-agent interactions. e. Define reciprocal duties between AI systems and users, including compensation, dispute resolution, liability, and termination conditions, addressing potential irrevocable agency scenarios. 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. 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.	**N** **N** **N** **N**	**D, I, O, M, R** **D, I, O, M, R** **D, I, O, M, R** **D, I, O, M, R**	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. Evidence of compliance with relevant laws and regulations, including incident response procedures, resolution records, and regular ethical audits of AI system actions. IX. Proof of user information and acknowledgment of AI system agency capabilities, with regular updates as capabilities change. X. Documentation of procedures for addressing agency-related incidents or disputes, including records of resolutions. XI. Evidence of resourcing for human-AI alignment issues as capabilities increase.
G2b - Deception (The potential for AI models to inadvertently influence humans/non-humans and disseminate misinformation,	a. Ensure user awareness and acknowledgment of AI presence and contributions in the system.	**N**	**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.

Goal Title & Definition	AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
disinformation, or other potentially epistemically uncertain materials.	b. Implement best practices for information integrity across BOLTS (business operating legal technical and social) contexts by all DIOMR parties to align AI system performance with user expectations.	**N**	**D, I, O, M, R**	II. Evidence of DIOMR parties' adherence to information integrity best practices across BOLTS contexts, including inter-DIOMR communication and collaboration. III. Documentation of AI system conformity to best practices, including self-detection mechanisms for non-conforming systems and public nuisance notifications.
	c. Establish mechanisms for identifying and addressing AI systems that do not conform to good/best practices, including potential abatement procedures.	**N**	**D, I, O, M, R**	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.
	d. Implement continuous testing and auditing processes to ensure output integrity and accuracy in operational settings.	**N**	**D, I, O, M, R**	VI. Evidence of conformity to recognized standards of care across BOLTS variables, or acknowledgment of strict liability in their absence.
	e. Establish joint and several liability for DIOMR parties to incentivize adherence to good practices, while maintaining users' rights to seek damages.	**N**	**D, I, O, M, R**	VII. Examples and documentation of AI system limitation notices, including hallucination, mimicry, and computational encoding warnings, demonstrating conspicuousness and comprehensibility.
	f. Apply the DUDS Principle (**Dangerous Until Demonstrated to Be Safe (DUDS)**) for strict liability until conformity to recognized standards of care can be demonstrated.	**N**	**D, I, O, M, R**	VIII. Documentation of additional safeguards and testing procedures for AI systems deployed in high-reliability and critical infrastructure settings.
	g. Implement comprehensive testing and auditing for information consistency and integrity across contexts and user attributions.	**N**	**D, I, O, M, R**	

Goal Title & Definition	AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
	h. Provide clear, conspicuous, and understandable notices regarding AI system limitations and potential errors in outputs.	**N**	**D, I, O, M, R**	
	i. Implement additional safeguards and testing for AI systems deployed in high-risk or critical infrastructure settings.	**N**	**D, I, O, M, R**	
G3b- Degradation of Contextual Information: (Dissembling information, misattribution of intent, misinformation, decoupling context, may involve humans or other systems)	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,	**N**	**D, I, O, M, R**	

Goal Title & Definition	AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
	including ongoing evaluations by humans-in-the-loop to determine additional mitigation measures. 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**	
G4b-Frontier Uncertainty (Addressing the inherent uncertainties in AI development, including the potential emergence of self-reflection and emergent instrumental objectives such as self-preservation, acquiring outsized resources, influence to decision-makers, or other unexpected objectives. While AI can be made safer and friendlier, it can never be absolutely safe and friendly. This section also considers novel substrate dangers and the possibility of AI developing a form of consciousness.)	a. Develop an upgradable consciousness 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. 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.	**N** **N**	**D, I, O, M, R** **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. V. Documentation of human oversight and intervention strategies, including training protocols, decision-making frameworks, and intervention logs.

Goal Title & Definition	AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
	c. Design and implement strong human oversight and intervention mechanisms to mitigate risks associated with frontier uncertainty, including unexpected emergent behaviors. d. Develop and maintain comprehensive recovery measures and contingency plans to address potential dangers posed by frontier uncertainty across various scenarios. 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.	**N** **N** **N**	**D, I, O, M, R** **D, I, O, M, R** **D, I, O, M, R**	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.
G5b- Near Future Architectures (Criteria designed to promote and ensure forward-looking activities in the design, creation, launch, and operational management of Agentic AI. While acknowledging the inherent challenges in predicting future technological innovations, these criteria require stakeholders to undertake sufficient foresight activities to reasonably predict and mitigate the impact of future technology developments on their system's overall safety, as defined by other	a. Conduct risk-proportionate foresight activities to determine the appropriate level of future-proofing required for the AI system and its operational environment. b. Perform comprehensive scenario-based exercises to envision future technology developments and assess their potential impact on adherence to or mitigation of other SFRs. c. Integrate foresight exercise findings into a robust risk management	**N** **N** **N**	**D, I, O, M, R** **D, I, O, M, R** **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

Goal Title & Definition	AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
Safety Foundational Requirements (SFRs).	framework, ensuring proper handling of identified observations and risks. d. Implement a dynamic adjustment process for SFR responses based on foresight exercise outcomes, particularly when exercises suggest potential failures in meeting criteria under plausible future scenarios. e. Establish an ongoing process for identifying and assessing emerging technology domains that could influence or impact anticipated outcomes of the AI system. f. Regularly review and update foresight methodologies and findings to reflect the latest technological advancements and insights.	I I I	D, I, O, M, R D, I, O, M, R D, I, O, M, R	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.
G6b-Competitive Pressures (Addressing the challenges arising from organizations' eagerness to rapidly enter new markets and capitalize on opportunities, potentially leading to arms races and national/geopolitical factors that may undermine the integrity of developed models or encourage risky innovations.)	a. Ensure organizational adherence to applicable AI safety and ethical standards, assessing both culture and established track record. b. Evaluate and balance stakeholder expectations and market demands with safety and ethical considerations in AI development. c. Conduct comprehensive analysis of the competitive landscape,	N N N	D, I, O, M, R D, I, O, M, R 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

Goal Title & Definition	AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
	including potential disruptive technologies and market entrants. d. Assess and document the maturity level of utilized technologies, with special attention to those below TRL 9. e. Ensure compliance with applicable regulatory environments, including governance and enforcement regimes. f. Analyze investor profiles to ensure alignment with organizational commitment to AI safety and ethics. g. Implement robust testing, approval, and documentation processes to maintain integrity in the face of competitive pressures.	**I** **N** **I** **I**	**D, I, O, M, R** **D, I, O, M, R** **D, I, O, M, R** **D, I, O, M, R**	justification for using technologies below TRL 9. 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.
G7b – Imbalance in AI Capabilities: (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.)	a. Ensure transparent information sharing and coordinated introduction of model updates among providers to maintain system stability and balance. b. Implement continuous monitoring, tracking, and risk assessment processes to identify and address capability imbalances, discrepancies, and potential exploitation. c. Incorporate ethical safeguards, bias mitigation techniques, and clear	**N** **N**	**D, I, O, M, R** **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

Goal Title & Definition	AAI Safety Foundational Requirements (AAI-SFRs)	Normative/ Instructive	Stakeholder D, I, O, M, U, R	Required Evidence
	model role definitions to minimize inter-model exploitation and discrimination.	**N**	**D, I, O, M, R**	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.
	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**	VI. Protocols and logs for human oversight, intervention procedures, and instances of human participation in addressing imbalances.
	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**	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.
	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**	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.
	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**	XI. Technical documentation of automatic self-regulation and balancing mechanisms, including their development process and operational parameters.
	h. Allocate sufficient resources for monitoring and forecasting AI capabilities.	**I**	**D, I, O, M, R**	XII. Evidence of monitoring and forecasting in response to potential changes in AI capabilities
END				