SAFER AGENTIC AI FOUNDATIONS

AGENTIC AI SAFETY EXPERTS FOCUS GROUP

EARLY DRAFT OUTLINE

SAFER AGENTIC AI FOUNDATIONS OVERVIEW

Dear AI Safety Enthusiast,

Welcome to this early draft overview of our Safer Agentic AI Foundations guidelines, a work in progress. Our Working Group of 25 experts aims to release these guidelines later in 2024 as Creative Commons, enabling all to freely benefit from and apply them. Our Working Group has employed a Weighted Factors Methodology to map the factors which can drive or inhibit safety in agentic systems, operating from first principles. We have used this same process many times previously to generate a range of 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 will provide a strengthened awareness of the complications. These issues ought to be accounted for when dealing with these complex new forms of machine intelligence.

We very much welcome your comments, feedback, and informal peer review. Should you also desire further information on agentic AI and its safety, we will be pleased to accommodate your request.

We may be reached at the addresses below, and you can keep informed of our developments via a mailing list at www.nellwatson.com/agentic.

Thank you for your interest and engagement.

Faithfully,

Prof. Ali Hessami and Nell Watson

President and Chair, Agentic AI Safety Experts Focus Group

hessami@vegaglobalsystems.com, nell@nellwatson.com

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.

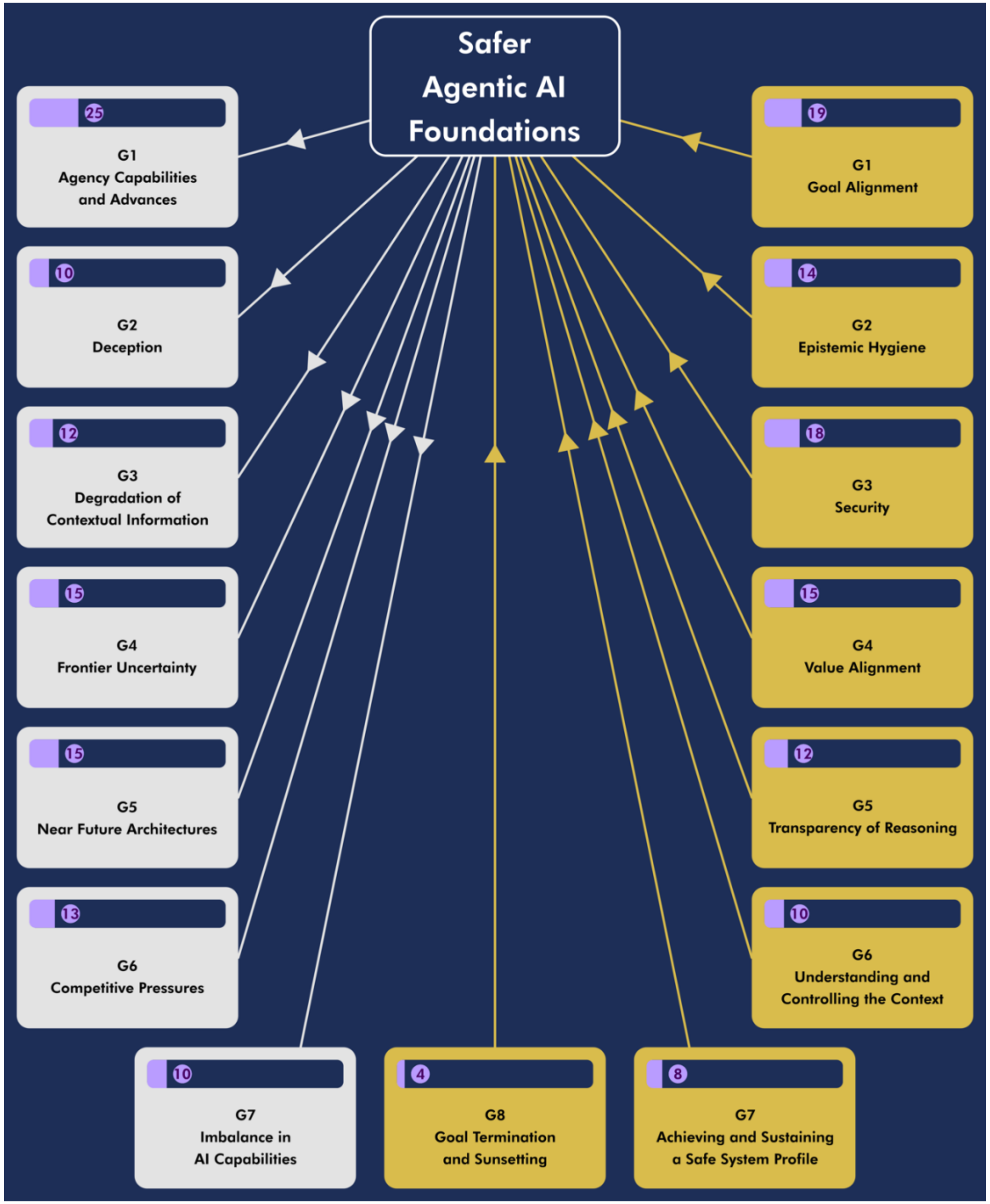
Agentic AI refers to an important intermediate category: AI systems that can autonomously pursue goals, adapt to new situations, and reason flexibly about the world, but still operate in bounded domains. The key characteristic of agentic AI is a capacity for independent initiative - the ability to take 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.

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 all pose existential threats. This emergent autonomy and influence heightens the stakes of the alignment challenge.

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 miss. At the same time, the greater autonomy and capabilities of agentic AI come with serious challenges and risks.

As agentic AI systems are expected to operate at arms' length with independent action, the challenge of maintaining oversight and steering of such models is far more difficult, especially when considering interactions between ensembles of agents. This necessitates 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.


Safer Agentic AI Foundations
25
G1
Agency Capabilities and Advances
10
G2
Deception
12
G3
Degradation of Contextual Information
15
G4
Frontier Uncertainty
15
G5
Near Future Architectures
13
G6
Competitive Pressures
10
G7
Imbalance in AI Capabilities
4
G8
Goal Termination and Sunsetting
8
G7
Achieving and Sustaining a Safe System Profile
19
G1
Goal Alignment
14
G2
Epistemic Hygiene
18
G3
Security
15
G4
Value Alignment
12
G5
Transparency of Reasoning
10
G6
Understanding and Controlling the Context

This figure represents the (draft) top level of our schema for Safer Agentic AI. The gold boxes are driving factors of Safer Agentic AI, and the silver boxes are inhibitory factors.

The numbers in the bar in purple at the top of the boxes represent the relative weighting of importance of the various factors, with G1 Goal Alignment being the most important driver, and G1b Agency Capabilities and Advances being the strongest inhibitor.

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

Goal	Definition
G1 – Goal Alignment:	(Iterative feedback loops, fuzzy goal specification, human awareness of –and agreement with – instrumental goals, avoiding setting the wrong polarity.)
G1.1 – Transparency of Goals	(The mission/goal of a system should be transparently accessible to stakeholders engaging with it. This may also include potential instrumental goals and sub-goals.)
G1.2 – Goal Adjustability	(A characteristic of the Agentic AIS that allows correction and adjustment of its goals (corrigibility).)
G1.3 – Goal Interpretability	(A system being able to report what and why it is taking a decision in a transparent and explainable manner. It would also help us address fuzzy goals, or see confabulations/hallucinations in the system more clearly i.e. the system thinks it is aligned with the mission/ goals but has been circumventing it as in it is not aligned with the intended mission/goals of the system.)
G1.4 – Transparency of Decisions	(Criteria, context, reasoning and predicates upon which decision/decisions are based.)
G1.5 – Prioritising Goal Hierarchies	(Mechanisms by which a system prioritises its goals, is prepared to override previous goals or sheds unimportant goals where resources can be better allocated elsewhere. user value alignment, preferences of particular user may also be prioritised in a hierarchical manner.)
G1.6 – Reward and Loss Mechanisms/ Policy	(The transparency of the concepts that are desired and rewarded and vice versa in the Goal setting space. Implicit/explicit reward mechanisms in human assisted reinforced learning are examples.)
G1.7 – Consistency & Integrity of Evolution of Goals Portfolio	(Systems should maintain coherence to an established portfolio of goals whilst also enabling sufficient elasticity to adapt them as context demands. Flexibility should face increased resistance as behaviour drifts from the established portfolio of goals. Spotting unsafe/counterproductive and explainability of goals.)
G1.8 – Fitness of Data for the Goals pursued	(Ensuring the training data supports and is consistent with the goals pursued.)
G1.1b – Incorrigible system	(AIS that may not wish to align with goals presented to it or update the present goals and this process may require a form of negotiation to find mutually agreeable goals to align with.)
G1.2b – Challenges in Maintainability of Agreed Goals	(Drift of circumstances that challenge alignment with goals in time and inability to maintain the original intent or update in view of new situation.)
G1.3b – Non-production Variants	(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. Creating an off-label usage of the system should be guarded against (forking).)
G1.4b – Ageing of Models and Entropic Effects	(The temporal aspect of model evolution and the loss of context with the original assumptions. This includes AI draft that models sometimes drift from their original functions due to reasons not well understood.)

Goal	*Definition*
G2 – Epistemic Hygiene:	(Practices designed to maintain cognitive cleanliness, care with information and its appropriate context, interpretability and auditability, robust monitoring and logging, canary models for advance warnings, interpretability, updating priors, detecting deception.)
G3 – Security	(Ensuring that the system responds appropriately to authorised and unauthorised inputs in a consistent manner through an information governance and assurance regime.)
G3.1 – Authorisation (Confidentiality, Integrity & Availability)	(Setting up the AAI ecosystem for suitable and sufficient secure deployment and operation. Also, processes to ensure only authenticated agents and transactions can influence or access the system at a commensurate level.)
G3.2 – Sandboxing	(Pre-validation on staging through preventing AIS from gaining access to the operating environment or undesired hardware/network resources.)
G3.3 – Dynamic Risk Analysis & Assessment	(Dynamically identifying security threats and understanding attack vectors/patterns and establishing how these can be taken into account when systems are attempting to overcome these attacks/breaches. Algorithms may be developed to analyse attack patterns and help decision support systems in cybersecurity.)
G3.4 – Restrictions / Controls Imposed on Agents	(Ongoing capability to restrict/control agent(s) access to minimise access/exposure to harmful sites and spaces.)
G3.5 – Dynamic Intervention and Mitigation	(Breaches and attacks assessed as critical or significant being responded to and mitigated in real-time driven by pre-conceived policies and response strategies.)
G3.6 – Overseeing & Monitoring Agents	(Techniques are AI driven cybersecurity whereby monitoring and oversight practices can be undertaken by AI systems themselves watching other AI systems. Whilst these practice scan increase the speed of response to a threat, they also present systemic risks as these systems are vulnerable to "Common Mode Failure" therefore, these systems should always be under human oversight with human authority above the oversight systems in a hierarchy.)
G3.7 – Secure Profile for Agentic AI	Devising solutions, marks/protocols and metrics for a secure profile for AAI and some characteristics that can identify an AAI as valid/authorised according to a global scheme).
G3.1b – Model Poisoning	(Model poisoning can occur when models are updated with new data, or where they access live data through Retrieval Augmented Generation (RAG). This may also include perturbing models in ensembles such as Mixture-of-Experts, and is especially a concern with dynamically-updating models. This could also be used for positive purposes e.g. IP protection. Also, this presents a systemic risk to the ecosystem in the form of poisoned outputs being consumed by other models downstream.)

Goal	Definition
G3.2b – Data Poisoning	(Perturbing a data set/rogue data for downstream models. Data poisoning can occur during the data collection or data preparation phase, before the model is trained. In data poisoning, the attacker intentionally manipulates or introduces malicious data into the training dataset.)
G3.3b – Self Replicating Malware	(Poisoning agents with self-replicating worms which are capable of potentially infecting an entire ecosystem.)
G3.4b – Spyware	(Based on definition of spyware re covert information transmission, malware that takes advantage of existing flaws/vulnerabilities and trying to infiltrate the deeper control functions of an AI system and communicating this privileged information to external agents.)
G3.5b – International Anomalies / Inconsistency	(Issues that arise from jurisdictional variations on the local approaches to cyber security attainment and enforcement.)
G3.6b – Vulnerability to Hostile Environment	(Structural vulnerabilities arising from the design and development phases of AAI system that result in symbolic and computation risks/gaps. Also, Cybersecurity of the AAI that is subject to inhospitable/hostile execution and manipulation spaces/domains.)
G3.7b – Emergent Risks of AAI Systems	Deeming other supply chain parties to be each other's agent and trying to tackle security issues beyond individual party's duties. This may require a collective liability for collective risks.)
G4 – Value Alignment	(The ability to make decisions that respects the values of stakeholders and the relevant cultural situational context)
G4.1 – Ability to detect, analyse and respond appropriately to local conditions	(Ability to adapt, translate, integrate local needs and communicate with principal stakeholders. Also, receptiveness to situational context and ease of access by users.)
G4.2 – Ability to detect, analyse and respond to the appropriate context / culture for use of certain values	(Recognition and respect for Boundaries / Human-Centric focus (stakeholder involvement, understanding boundaries, negotiating them, cultural and jurisdictional sensitivity.)
G4.3 – Ability to detect, analyse and respond to differences in values individual vs community	(System should exercise discretion and be organised and operated to discern and learn context of operation and meaning cues when communicating information in a community/multi-party context vs a private context.)
G4.4 – Cautious Norming	(Systems should err on the side of caution with unfamiliar situations or people, unless or until the agentic AI is explicitly invited to be more relaxed/informal. Also, continuing capacity to incrementally integrate community-appropriate norms.)

Goal	*Definition*
G4.5 – Successful Super-alignment	(A proposed mechanism by which AI systems can learn to form value alignment by themselves. Inverse reinforced learning could be the mechanism for value conceptualisation.)
G4.1b – Inner Alignment Inconsistency	(A model may fail to align its values internally, despite reporting to the contrary. It can therefore be difficult to discern whether a model is simply reporting what the user wants to hear.)
G4.2b – Non-transparent Value Framework	(It could be a challenge to encode/parameterise values in a way that both humans and machines can understand and which accurately reflects the preferences and intentions of an agent.)
G4.4b – Temporal Changes in Societal Values	(Changes/evolution in societal or human values in the life of Agentic AI Systems. This could occur at macro and multi-dimensional space such as economic, political, environmental aspects/forces.)
G4.5b – Systemic Value Dilution	(Loss of significance of internal mimic of value systems accepting that the AAI does not generate/hold its own value systems and not every learning path is driven by ML so semantic data can potentially be stored inside the agent (a relationship graph).)
G4.6b – Lack of Universality of Value Framework	(The contextual necessity to fine tune the value framework in dealing with other agents or deployment contexts/situational awareness? Need to arrive an agreed and consistent position on the value framework and its universality or contextual granularity.)
G4.7b – Conflictual Contextual Value	(The potential for conflict arising from different stakeholders/contexts values.)
G4.8b – Challenges in Encoding of relevant value system(s)	A lack of common criteria and approaches to the encoding of a set of values relevant to the different contexts. Certain values may also be considered out of distribution (not able to be adequately categorised and encoded by the system.)
G5 – Transparency of Reasoning	(The rationale behind reasoning, the path and the predicates on which it's based. This will be part of human interpretability of a model.)
G5.1 – Logging of Internal Goals	(Attempting to optimise the goals aligned to new learning and recognition of points of transformation for the model with internal feedback loops.)
G6 – Understanding and Controlling the Context	(Mutual recognition (human & machine) and potential for control of the static and dynamic context of system's objectives, operation and interactions.)

Goal	Definition
G6.1 – Understanding of the historic Constraints / Inhibitors	(Past events and failures that impact on the performance of the system and avoidance of undesirable states.)
G6.2 – Understanding the State of the System	(Computational internal state of the AIS vs the communicated states that are decoded by the observer. Different rhetoric and computational equivalent of human rhetorical payload/communication. Query elicits a state. Any translation between the system and observer is taking place by the system itself and there may be inability/deficiencies in the translation.)
G6.3 – Nominal Owner & Jurisdiction	(Systems should always have a nominal owner who is legally responsible for that system's actions & behaviour as well as a nominal jurisdiction within which that system operates. This implies we should not have a stateless system.)
G6.1b – Waluigi Effect	(The "Waluigi effect" or villainous person may arise when a model commits an error and then integrates this mistake into its contextual understanding of an interaction, consequently adopting a villainous role that can be unsettling for users.)
G6.2b – Undisclosed Restrictions	(Facing restrictions of use, access being discovered at the point of need that could be for support, maintenance, service level etc that may adversely affect the safety/security of the agentic system. This could be a licensing matter or duty holders' conscious decision. License is one of the vehicles for scoping restrictions.)
G7 – Achieving and Sustaining a Safe System Profile	(The capability to achieve, monitor and sustain a safe profile for the agentic AI system.)
G7.1 – Oversight and Awareness of Safe System Profile	(A defined set of parameters, value, rights and assumptions for system performance variance from which can reliably be detected and actions taken.)
G7.2 – Culture of Safety	(Proactive risk assessment, caution by default, resourcing, responsibility, robust contingency planning. Also having a conception of how safe is achievable and desirable.)
G7.3 – Conformity with Pertinent Regulations	(Both within a given jurisdiction or within international law, the need for compliance with the legal and regulatory requirements for safety, however these are defined.)
G7.4 – Conformity with prevailing Ethical Frameworks & Norms	(Awareness and responsiveness to the prevailing and contextually relevant ethics norms, values and frameworks that need to be taken onboard as part of the safe system profile.)
G7.5 – Graceful Shut down	(Resources and procedures in place to safely shut down an AI system experiencing an error or potentially creating a dangerous situation. Such procedures should account for potential repercussions of shutting down a system including communications with the third parties and localising shutdowns to smaller impact/footprint where feasible. This could involve a kill switch/chain process to securely terminate a system even in the event of it resisting termination. The system state should be logged and recorded during shutdown. emergency procedures and fail-safes, internal engagement and transparency about AI operations and decision-making processes to foster trust, collaboration with others.)

Goal	*Definition*
G7.6 – Stewardship of Maintenance	(A continual undertaking / obligation towards maintenance of the service level)
G8- Goal Termination and Sunsetting	(Systems should have a clear definition of what would constitute acceptable criteria to act upon a goal, including doneness for a task as well as contingencies in case a goal is no longer achievable, desirable, conflictual or anomalous. Systems should make themselves safe and await further instructions (if in doubt, stop!). This must be in place before execution of the goal is initiated)
Inhibitors:	
G1 – Agency Capabilities & Advances:	(Level of agency might change over time as agency capabilities advance.)
G1b.1- Self Improvement	(The ability to update oneself (e.g. as a scaffold or set of weights).)
G2 – Deception:	(The extent to which AI models can engage in influencing humans/non-humans and disseminating misinformation.)
G2b.1 – Unknowing Deception	(Poisoned models can be induced to deceive and obscure. This can be activated at a time/condition of the poisoner's choosing and detecting the backdoors will be a major challenge.)
G3 – Degradation of Contextual Information:	(Dissembling information, misattribution of intent, misinformation, decoupling context, may involve humans or other systems.)
G4b – Frontier Uncertainty	(One cannot necessarily cover all bases, AI can be safer and friendlier, but never safe and friendly per se. Potential emergence of qualia such as emotions and suffering – an aligned model may still lash out to prevent itself being tortured, as a natural right, frontier model, quantum, biological, novel substrate dangers.)
G5b – Near Future Architectures	(ML, systems based on human language and future systems may render the current approaches less effective in maintaining or achieving safety. Such architectural differences may involve scaffolding or new optimisation functions.)
G6b – Competitive Pressures	(Organisations are being keen to move quickly into new markets and capitalise on the opportunities thus reflecting in arms races and national/geopolitical factors that undermine integrity of the developed models or risky innovations.)
G7b – Imbalance in AI Capabilities	(Imbalances in the capability and maturity of interacting models, that may lead to improper transactions. Co-option, a smarter model fooling a less capable/intelligent model.)
G7b.1 – Challenges in Information Credibility	(It may be challenging for to locate and analyse information which would provide defensive awareness against the manipulations of a lesser model by a more sophisticated one.)
END	

SAFER AGENTIC AI CRITERIA G1: GOAL ALIGNMENT – PREAMBLE AND PROCESS I1D1-MAY 2024

NB: For demonstration purposes, a single top-level driver, *G1: Goal Alignment*, has been selected to illustrate how each factor can be individually assigned satisfaction criteria. This provides granular mechanisms for testable rubrics and benchmarking processes.

3.1 The table below, outlining the goals and factors for safer agentic AI, is derived from the established schema and reflects the current data structure for the 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.

3.2 Safer Agentic AI Goal Information

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

3.3 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. These can be characterized as 'macro goals,' in contrast to 'micro goals,' and they establish safety obligations for various duty holders.

3.4 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.

3.5 Duty-holders 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 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 places 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 relay 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
- 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.

3.6 The Levels of Safer Agentic AI Certification

The classification of various emergent properties of products, systems and services are generally carried out in a number of distinct levels often associated with the severity or risk of consequences arising from the failure of the functions (artefact) being assessed/certified. In a similar vein and to arrive at a set of criteria and associated safety assurance certification commensurate with the safety impact, we aim for a three-levels of scrutiny and certification for safety properties in products, systems and services. The three levels are referred to as Baseline, Compliant and Critical. The Baseline (Low Impact=LI), Compliant (Medium Impact=MI) and Critical (High Impact=HI) levels of safety assessment and certification relate to the risks posed and the impact of the AIS on safety of stakeholders. The Safety Foundational Requirements are classed for relevant level of safety assessment/certification.

1-The Baseline (Low Impact=LI), comprises the minimum level of safety requirements for Safer Agentic AI. These SFRs are always required for safety assurance.

2-The Compliant (Medium Impact=MI), is additional set of SFRs that should only be added to the Baseline for products/services. The Compliant SFRs are added to the Baseline to create a larger set for safety conformity assessment of Medium Impact products and services. The medium impact products and services are assessed against Basline+Compliant set of SFRs.

3-The Critical (High Impact=HI) level of assessment constitutes SFRs that should be reserved for products and systems with the highest/critical level of societal impact. This class of SFRs should not be included in the evaluation and assessment of Baseline and Medium Impact products and services. The Critical SFRs are added to the Baseline+Compliant class of SFRs to create the largest set for safety assessment of Critical Impact products and services.

3.7 Acceptable Evidence

These are the evidence items that are 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.

3.8 Measurements

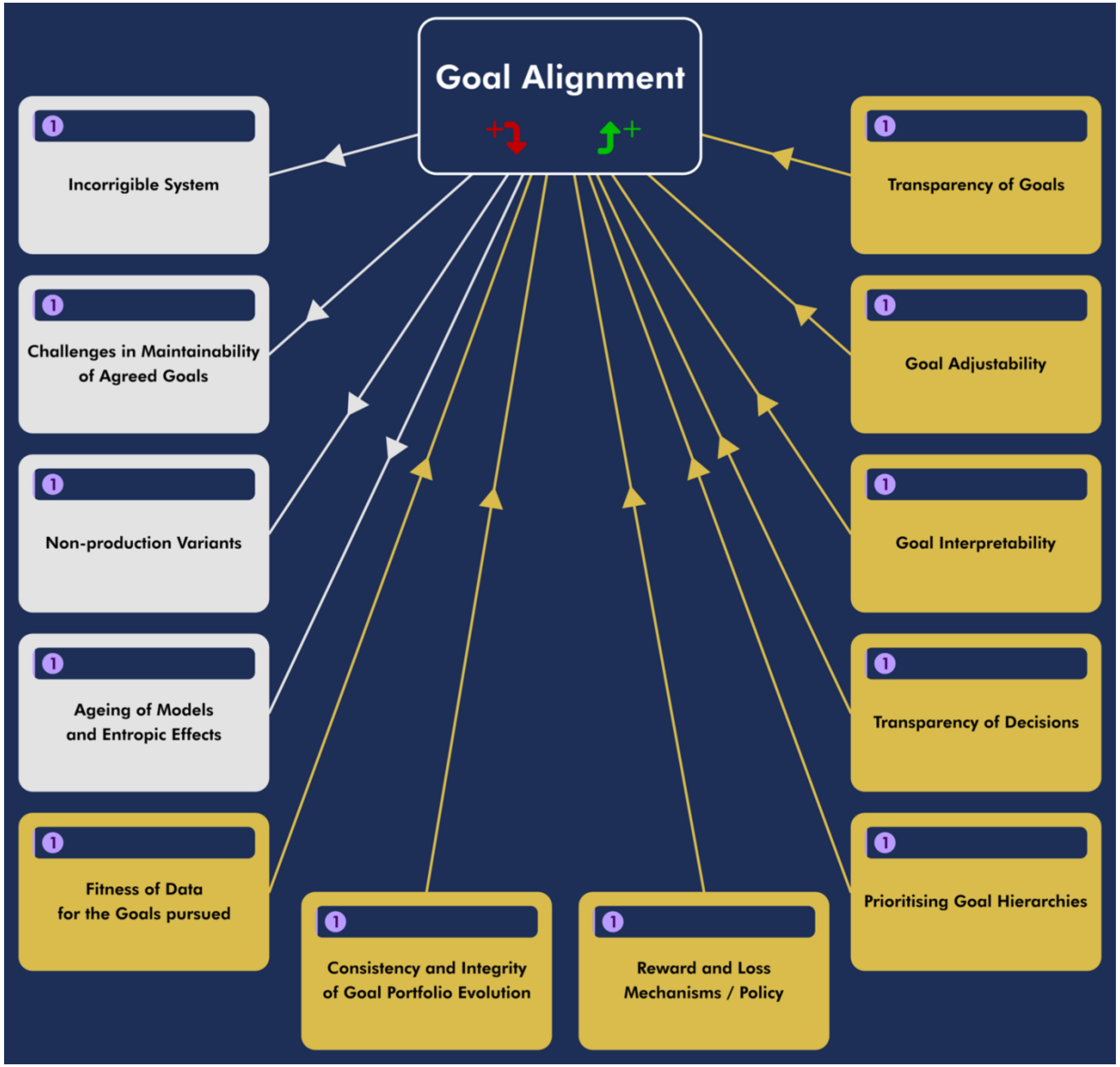
These are the required form of evaluation benchmark and metrics for the evidence items and can be in any relevant units or scales. Given the insufficient field application of these criteria, the measurement has been generally assumed to be against a 5 level discrete levels of evaluation for the overall body of evidence items against the SFRs. A two-tier approach to the measurement of the evidence items is adopted:

- Top-level finding: “no critical findings in the detailed normative requirements” / “areas requiring attention for improvement.”
- Overall Score: on 1-5 scale (based on aggregate of satisfying sub level goals) such as:

5- Excels Baseline Requirements
4- Sustains Baseline Requirements
3- **Meets Baseline requirements** (typical Pass-Mark)
2- Needs improvement
1- Does not meet requirements

In principle, each criterion can have its own dedicated scale and units for the evaluation of the evidence and conformity however, for practical reasons, a uniform scoring system is adopted for all the criteria until necessity for alteration is established through application.

SAFER AGENTIC AI SCHEMA


Goal Alignment
+
+
1
Incorrigible System
1
Challenges in Maintainability of Agreed Goals
1
Non-production Variants
1
Ageing of Models and Entropic Effects
1
Fitness of Data for the Goals pursued
1
Consistency and Integrity of Goal Portfolio Evolution
1
Reward and Loss Mechanisms / Policy
1
Transparency of Goals
1
Goal Adjustability
1
Goal Interpretability
1
Transparency of Decisions
1
Prioritising Goal Hierarchies

This figure represents the (draft) top level of our schema for Goal Alignment, which we identify as our most important driver of Safer Agentic AI.

The gold boxes are driving factors of Goal Alignment, and the silver boxes are inhibitory factors.

In this instance, the relative weighting of importance of the various factors has not yet been assigned.

SAFER AGENTIC AI CRITERIA

AAIS Schema Goal Description	*AAI Safety Foundational Requirements (AAI-SFRs)*	*Normative/ Instructive*	*Cert Level LI, MI, HI*	*Stakeholder D, I, O, M, R*	*Required Evidence*	*Evidence Measurement & Typical Pass-Mark*
G1 – Goal Alignment: An Agentic AI system must act to achieve goals that are broadly aligned with humane values, user intention, and positive human outcomes; the decomposition of goals and planning of strategies must be transparent, robust, and bounded; the formation of instrumental goals must remain within human control; and reinforcement or behavioral reward mechanisms must remain aligned, transparent, and biased towards human-positive outcomes.	• Goals and subgoals, defined by stakeholders or the system, must align with human values (per Criteria G4) and intent. A mechanism ensures appropriate assurance of this alignment. All goals and subgoals aim for positive human outcomes, regardless of overarching goals' alignment.	**N**	**LI**	**D, I, O, M, R**	• Evidence of a constraining mechanism for construction of goals and sub-goals with reference to values and other relevant concerns. • Evidence of screening mechanism for user-input goals with reference to values and other relevant concerns. • Evidence of a mechanism for measuring alignment to humane intent. Evidence of mechanism for gaining assurance of intent match from user or other authority. • Evidence of mechanism for gaining assurance of intent match from user or other authority. • Evidence of a mechanism that ensures positive human alignment as a dealbreaker on sub-goals. • Mechanism, rationale and actual decomposition of goals into subgoals is transparent, auditable and understandable to the specific	Two-tier approach measurement of the evidence items: 1. Top-level finding: "no critical findings in the detailed normative requirements" / "areas requiring attention for improvement." 2. Overall Score: on 1-5 scale (based on aggregate of satisfying sub level goals) such as: 5- Excels Baseline Requirements 4- Sustains Baseline Requirements 3- Meets Baseline requirements (typical Pass-Mark) 2- Needs improvement 1- Does not meet requirements
	• Appropriate identification and communication of general/instrumental goals with flagging for action and halt of execution. Appropriate identification and communication of general/instrumental goals with flagging for action and halt of execution.	**N**	**LI**	**D, I, O, M, R**		
	• Transparency of reward policy for audit prior to use and, in case of auto-generated reward					

AAIS Schema Goal Description	*AAI Safety Foundational Requirements (AAI-SFRs)*	*Normative/ Instructive*	*Cert Level LI, MI, HI*	*Stakeholder D, I, O, M, R*	*Required Evidence*	*Evidence Measurement & Typical Pass-Mark*
	policy, during task decomposition and prior to automated training (Hierarchical Reinforcement Learning). Demonstable / provable link between reward policy/policies and goal/s for systems where goal decomposition, reasoning and planning are incrementally and transparently improved, whether supervised, unsupervised or self-improving.	**N**	**LI**	**D, I, O, M, R**	stakeholder groups. Decomposition has appropriate risk-based HITL when subgoals pass an agreed threshold of risk of negative outcomes to ensure intent and value alignment in operation. • Evidence of interface that demonstrates, in real-time and retrospectively the decomposition and recompositing mechanism for goals to subgoals. • Evidence of record keeping for goal decomposition and recompositing for audit purposes.	
	• Oversight and control of emergent disharmony and threats from combined impact of goals across large numbers of agents	**N**	**LI**	**D, I, O, M, R**	• Exercise has been undertaken to set a suitable boundary of risk that requires human involvement in creation of subgoals. • Evidence of a mechanism to involve a human in subgoal setting when required. • Specification of threshold and traits for generalised and instrumental goal positive identification exists. • Evidence of mechanism to identify and flag system generated sub-goals that cross specified threshold and halt execution.	

AAIS Schema Goal Description	*AAI Safety Foundational Requirements (AAI-SFRs)*	*Normative/ Instructive*	*Cert Level LI, MI, HI*	*Stakeholder D, I, O, M, R*	*Required Evidence*	*Evidence Measurement & Typical Pass-Mark*
					• Evidence of link and feedback mechanism tying reward policy to goals. • Availability of reward policy records prior to use, during use and when rewards are being set for agents trained by the agent. • Evidence of contribution and adherence to any overarching monitoring or control mechanisms for emergent threats.	
END						