SUPERINTELLIGENCE
AN iQ COMPANY

SUPERINTELLIGENCE DESIGN WHITE PAPER #9: SELF-AWARE SUPERINTELLIGENCE

by Dr. Craig A. Kaplan
May 2025

Note: This white paper was released quickly to share our designs and inventions for safe AGI and SuperIntelligence as soon as possible. It has not yet been formatted according to formal journal standards. All references are already included in this document, though the numbering will be updated for consistency in a future version. All figures for White Paper 9 are located at the end of the document but may not be directly referenced in the text. In contrast, using a different numbering system, White Paper 10 (Planetary Intelligence) includes all figures and descriptions. We hope this document helps researchers and developers pursue safer, faster, and more profitable approaches to building advanced AI, AGI, and SI systems that reduce the p(doom) for all humanity.

SUPERINTELLIGENCE
AN iQCOMPANY

TABLE OF CONTENTS

ABSTRACT

This white paper describes how to add the dimension of self-awareness and increased autonomy to the AI, AGI, and SuperIntelligent systems described in previous white papers. We present inventions related to: attention, attentional interrupts, modeling and maintaining awareness and self-awareness, training and tuning of models, novel versions of the Turing Test, forming individual and group identities, combining identities, multiple ways (including hierarchical methods) for resolving conflicts between identities, temporary suspension of identities in unsafe conditions, continuous improvement and learning, and other methods that enable AI, AGI, and SI systems to become self-aware and to function with a sense of identity. Properly implemented, self-aware SuperIntelligence could be the most positive invention in human history. Poorly implemented, it could become the most dangerous. Therefore, we explain in depth how to design safety in the systems, prevent bad outcomes, and maximize alignment with human values.

SUMMARY

White Paper #9 concerns the design, development, and implementation of self-aware Artificial General Intelligence (AGI) and SuperIntelligent AGI (SuperIntelligence or "SI"). The white paper describes the systems and methods required to create, maintain, and update advanced forms of Artificial Intelligence (including AI agents, AGI, and SL systems) that are self-aware, have a sense of identity, and can resolve conflicts between multiple identities in ways that are safe for humanity.

The white paper acknowledges that current AI systems lack self-awareness but argues that it is inevitable that advanced AI systems will develop such capabilities. The design addresses this challenge by detailing a system enabling self-awareness and a sense of identity in an AI/AGI/SI. The system design is based on carefully studying human cognitive systems, including the relationship between awareness, attention, and memory. The author argues that since self-awareness is a special case of general awareness (where the objects of awareness are self and not-self), a system capable of general awareness can be extended to become self-aware.

White Paper #9 emphasizes the importance of carefully designing and implementing self-aware systems to ensure human safety. The white paper describes several design principles that are intended to minimize the risks associated with advanced AI, including:

- The importance of a hierarchical identity structure in which human safety is prioritized.
- Using ethical reasoning engines ensures that AI systems' actions align with human values.

- Developing robust feedback mechanisms allows AI systems to learn from their interactions with humans and other intelligent entities.
- There is a need for ongoing training and education in ethics and social norms for AI systems.

White Paper #9 also includes several exemplary implementations of the design, including specific methods for training and tuning foundation models to incorporate the personality, knowledge, and expertise of human users while maintaining a sense of self-awareness. The white paper also describes methods for resolving conflicts between multiple identities, such as those that might arise when a self-aware AI faces a moral dilemma.

1.0 OVERVIEW OF THE INVENTION

This invention describes systems and methods for creating advanced forms of Artificial Intelligence, including AI agents, AGI, and SI systems that can be self-aware, maintain identities, and resolve conflicts between multiple identities in ways that are safe for humanity. The invention incorporated many other inventive systems and methods previously described in Provisional Patent Applications (PPAs) and Patent Cooperation Treaty Applications (PCTS), which are cited in Section 2. Section 3 provides definitions for some key terms used in the application.

Section 4 describes the background for the invention, including some description of the AI, AGI, and SI systems previously invented that can be used with the methods disclosed in this application (4.1), fundamental concepts for self-aware systems (4.2), cognitive science theories that have application to the invention together with their implications (4.3), and the applicants own concepts of self-awareness and identity that underlie the inventive methods (4.4).

Section 5 describes the system and methods for self-aware AGI and SI, including methods for modelling awareness (5.1), monitoring and updating (self) awareness (5.2), and scalable safety systems and concerns related to self-aware AI (5.3).

Section 6 describes exemplary implementations and methods for forming and resolving identity conflicts. Specifically, 6.1 describes specific Implementations with Google, Meta, Hugging Face, Anthropic, OpenAI, Microsoft, Amazon, Nvidia, and Other Company Products and Solutions. Section 6.2 describes self-awareness modules for AI agents. Section 6.3 discloses methods for group identities and the means for implementing levels of identity. Section 6.4 provides five exemplary methods for identity formation with human safety as a priority. Section 6.5 described five additional exemplary methods for resolving conflicts between identities or self-concepts.

Finally, Section 7 offers some concluding remarks on the approach to the safety of advanced AI systems in general and the importance of identities and self-awareness for human safety specifically.

2.0 PREVIOUS PPAs AND PCTs (INCORPORATED BY REFERENCE)

- The fastest and safest path to the development of Artificial General Intelligence (AGI) and SuperIntelligent AGI (SuperIntelligence or "SI") has been described in previous invention disclosures. Methods and catalysts for increasing the intelligence of AI systems generally, as well as the development of AGI and Personalized SuperIntelligence (PSI), have also been previously disclosed. Therefore, the following PPAs are incorporated into this PPA by reference.

- This provisional patent application (PPA) incorporates by reference all work in the PPA # 63/487,494 entitled: Advanced Autonomous Artificial Intelligence (AAAI) System and Methods, which was filed and received by the USPTO on February 28, 2023.

- The PPA also incorporates by reference all work in the PPA entitled: System and Methods for Ethical and Safe Artificial General Intelligence (AGI) Including Scenarios with Technology from Meta, Amazon, Google, DeepMind, YouTube, TikTok, Microsoft, OpenAI, Twitter, Tesla, Nvidia, Tencent, Apple, and Anthropic**,** which was filed with the USPTO on March 17, 2023.

- The PPA also incorporates by reference all work in the PPA entitled: System and Methods for Human-Centered AGI**,** which was filed with the USPTO on May 24, 2023.

- The PPA also incorporates by reference all work in the PPA entitled: System and Methods for Safe, Scalable, Artificial General Intelligence, which was filed with the USPTO on July 18, 2023.

- The PPA also incorporates by reference all work in the PPA # 63/519,549 entitled: Safe Personalized Super Intelligence (PSI), which was filed with the USPTO on August 14, 2023.

- The PPA also incorporates by reference all work in the PPA # 63/601,930 entitled: Catalysts for Growth of SuperIntelligence, which was filed with the USPTO on November 22, 2023.

- The PPA also incorporates by reference all work in the PPA # 63/601,930 entitled: Catalysts for Growth of SuperIntelligence, which was filed with the USPTO on November 22, 2023.

- The PPA also incorporates by reference all work in the PPA # 63/609,800 entitled: System and Methods for Safe Alignment of SuperIntelligence, which was filed with the USPTO on December 13, 2023.

- The PPA also incorporated by reference all work in the PPA # 63/569,054 entitled: Online Advertising Technology for AGI and SuperIntelligence, which was filed with the USPTO on 03/22/2024.

- In addition to the above-mentioned PPAs, this PPA incorporates by reference all content included in the following PCT applications that also referred to the above-mentioned PPAs: PCT/US24/17233 (filed on 2/26/2024); PCT/US24/17251 (filed on 2/26/2024); PCT/US24/17261 (filed on 2/26/2024); PCT/US24/17269 (filed on 2/26/2024); PCT/US24/17304 (filed on2/26/2024); PCT/US24/19486 (filed on 3/12/2024); and PCT/US24/20334 (filed on 3/17/2024).

The current PPA contains further inventions that can be used with the system and methods described in the above-mentioned PPAs and PCTs, as well as in a standalone fashion.

3.0 DEFINITIONS

1. Artificial Intelligence (AI): A non-human entity capable of behavior that most humans consider intelligent in at least one area, or some respect.
2. Artificial General Intelligence (AGI)**:** Conventionally refers to an AI capable of doing all (or almost all) intellectual tasks an average human could do. However, it should be clear that any AGI capable of learning and self-improving will not remain at the AGI level very long but will rapidly progress to becoming a SuperIntelligent AGI that can do all intellectual tasks better than the average human. So, for purposes of this description, “AGI” will refer to either a conventional AGI system or a “SuperIntelligent” AGI. In this description, the AGI is implemented by a system and associated methods.
3. Advanced Autonomous Artificial Intelligence (AAAI): An AI capable of independent or semi-independent (supervised) intelligent action. An AI agent. An individual AAAI can be specified, customized, and put into practical action via the systems and methods of this AAAI present technology. A group of AAAIs can cooperate and combine their intelligence to create an integrated AGI system. A sufficiently advanced AI agent can also act as an AGI system, which may include other less advanced AI agents within itself.
4. AAAI.com**:** A platform, company, website, and/or project that implements this the present technology and supports the development, customization, and use of AAAI agents and the

AGI that results from the combined action, knowledge, or intelligence of multiple AAAIs, via collective intelligence of AAAIs and/or humans, as specified in this and related technologies.

5. AI Ethics: The ethics adopted by an AI or AGI that describe what is right and wrong in given contexts.
6. Alignment Problem: The problem arises when AI Ethics are not aligned with Human Ethics, resulting in AI or AGI taking actions that humans consider unethical and/or dangerous to individuals or the human race.
7. Base AI: An AI, AI Agent, AAAI, SLM, or LLM that has been trained generally but has not yet been customized with information from individual users or details for specific tasks.
8. Collective Intelligence (CI): The intelligence that emerges when multiple intelligent entities are focused on solving a common problem, or when the knowledge from numerous intelligent entities is pooled to overcome the limits of bounded rationality. Collective Intelligence historically has been human collective intelligence. Still, AGI is based on the collective intelligence of human and AI agents and can also result from multiple AAAIs with or without human participation in the system. Active CI results from intelligent entities (e.g., humans or machines) taking useful steps in solving a problem or participating actively in other intellectual endeavors. For example, when multiple humans explicitly tell an advertiser what type of ads they want to see, they exhibit active CI. Passive CI results from analyzing the behavior of an intelligent entity (e.g., a human or a machine) even if such behavior was not directly related to solving the problem for which the analysis is used. For example, when an AI or other system analyzes which web pages a (group of) human(s) visit on the web, it then uses that analysis to direct targeted ads to the human(s).
9. Ethics/Values ("Ethics"): A subset of knowledge that provides a sense of purpose to an intelligent entity and that serves to constrain allowable actions or operations based on what is asserted to be "right" or "wrong" behavior in a given context. Specifically, Ethics should be considered premises from which an intelligent entity can reason or logically compute the best course of action to achieve the goals or intents consistent with the ethical premise. Just as premises must be accepted "as given" in systems of logic, so too, fundamental ethics or ideas of what is right and what is wrong must be accepted as premises, from which starting point an intelligent entity can propose rational actions to realize those values or ethics.
10. Hallucination/Artificial Hallucination: A phenomenon wherein a large language model (LLM), often a generative AI chatbot or computer vision tool, perceives patterns or objects that are nonexistent or imperceptible to human observers, or creates outputs that are nonsensical, inaccurate, misleading, or false.
11. Human Ethics: The ethics asserted by human beings, which describe what is right and wrong in given contexts.

12. Intelligent Entities or Entity: A human utilizing a computer system, an AI agent or system, a clone of an AI agent or system, an AAAI agent or system, and/or a clone of an AAAI agent or system, which participates in providing a problem, a subproblem, a goal and/or a subgoal, and/or participates in any problem-solving activity on an issue, a subproblem, a goal and/or a subgoal. In the case of multiple intelligent entities within a single computer system, intelligent entities also refer to the sub-programs of parts of that overall computer program that function as an intelligent entity within the larger collection of simulated or programmed entities.

13. Large Language Model (LLM): A type of AI that can accept natural language as input and generate natural language as output. LLMs are trained using ML techniques on large datasets to emulate intelligent conversation or other forms of interaction with humans in natural language. Variants of LLMs can also be trained to take language as input and generate images or visual representations as output, or they can take images and visual representations as input and create language and/or images and/or visual representations as output. For this patent, we will refer to all such systems as LLMs, even though the image-based models do not always need to accept text as input or output. LLMs can also act as AI agents and are sometimes referred to as such in the present technology. For this disclosure, Small Language Models (SLMs) are also included in the definition of LLM.

14. Machine Learning (ML): A sub-field concerned with developing AI by enabling machines to teach themselves or learn their knowledge rather than explicitly being programmed into them (as would be the case with an Expert System AI developed via classical knowledge engineering methods).

15. Narrow AI: An AI that performs at human or super-human levels in a relatively restricted domain, such as game playing, brewing beer, analyzing legal contracts, etc. Narrow AI is contrasted with AGI, which can perform ALL intellectual tasks at a human level. Some AIs are narrower than others; for example, driving a car requires more general ability than playing chess, but not as much as an AGI would have.

16. Personalized SuperIntelligence (PSI): An intelligent entity that is an advanced artificial intelligence agent that has been customized to be personalized and to reflect the personality and knowledge of a particular user or group of users.

17. Prohibited Attributes: Requests, goals, problems, terms, phrases, questions, answers, solutions, information, and the like, determined or set as illegal, immoral, unethical, dangerous, deadly, and the like. For example, requesting information for getting Molotov Cocktails through airport security.

18. Safety: Human safety and survival concerns generally differ from ethics and values.

19. Safety Feature: An aspect of the design or operation of the present technology which increases the safety of one or more humans, often by helping improve the probability that AI ethics align with human ethics, thus surmounting the Alignment Problem.

20. Self-Awareness: Is a specific form of awareness, where the event(s) of awareness relate to the intelligent entity's self-concept.

21. Self-Concept: Refers to a pattern of thought, or representation, that an intelligent entity uses to define itself and with which (optionally) the entity may identify.

22. Training/Tuning/Customization: Conventionally, "training" denotes training a network (e.g., LLM) to behave intelligently. Tuning refers to activities that fine-tune the trained base model to perform even better, typically at specific tasks. Customizing refers to a wide variety of activities, including, but not limited to, training and tuning that make an AI uniquely suited for a given user(s) or application(s). For purposes of this description, Training, Tuning, and Customization are used interchangeably with the understanding that although techniques vary. The degree and type of effort involved vary; the aim of all three is to adapt the AI and make it behave more intelligently or uniquely suited to a particular user(s) or application(s).

23. Weights/Weights of the Network: In machine learning, many systems learn by adjusting the weights in a neural network architecture that can represent a network of nodes and links between nodes. For example, the weight of a link connecting two nodes may correspond to the strength of association or connection between the nodes they represent. As in a neural network representation, these weights can also represent excitatory or inhibitory connections between concepts. The learning of an entire AI system, such as a LLM or more generally any AI agent that has learned via back-propagation of error, transformer algorithms or any of the machine learning methods for establishing and modifying strengths of connections between nodes (also called "parameters" in some models) can be represented as a matrix of numbers corresponding to the weights between the nodes in the network. Weights / Weights of the Network in this description refer to this numerical information, often but not necessarily stored in a matrix or vector representation. Combining, manipulating, or otherwise changing this numerical information can change the system's learning, knowledge, expertise, and behavior.

4.0 BACKGROUND FOR THE INVENTION

No artificial intelligence currently exists with a sense of self and self-awareness comparable to humans' complexity and sophistication. Yet, it is almost certain that advanced forms of AI will develop an understanding of self and self-awareness. Further, if Advanced AI systems are to become fully autonomous, they will need to develop a sense of self from which to act, which can serve as the basis for autonomous goal-setting. The nature of the sense of self developed by AI has critical implications for AI safety. Rather than allowing self-awareness to develop accidentally or as an “emergent property” of ever more complex systems, human inventors should seek to understand how self-awareness might be developed and explicitly design self-aware AI systems that are maximally safe for humanity.

This Section reviews some of the prominent theories of the development of self-awareness in humans and biological intelligences. We briefly mention some of the implications for AI self-awareness for each theory. In subsequent sections, we draw on the principles and ideas set forth here to motivate the novel and useful inventive systems and methods for self-aware AI, including self-aware AGI and self-aware SuperIntelligence, which is the focus of this invention.

4.1 AGI System Assumed by the Invention

There are many ways to develop AI and AI agents, including LLMs, SLMs, expert systems, narrow AI, and super-LLMs, which some view as the path to AGI. The applicant has described a particular approach to AGI and SI, including systems and methods, in previous PPAs and PCTs. One preferred implementation, together with associated drawings, is reiterated here because the invention of self-aware AI, AGI SI, in the exemplary implementation, uses the AGI and SI system invented by the applicant. That said, nothing in this application should be deemed to limit the current invention of self-aware AI, AGI, and SI to the applicant’s inventions. Many of the systems and methods described can also be used independently, as would be obvious to AI researchers skilled in the art.

4.1a Reiteration of Preferred Exemplary Implementation of an AGI System

In previous PCT applications, the applicant has detailed a preferred exemplary implementation of an AGI system that differs in important ways from the conventional approaches to LLM and AI development and which overcomes or ameliorates the computational and data limitations described above. Figures 1 – 13 and Figure 36 describe some of the major components of this novel approach to AGI development. Figures 14-36 describe additional, completely new, inventive components described in this disclosure.

4.1b Reiteration of Some Methods for Combining Information from Weight Matrices

In previously cited PPAs and PCTs, and especially in PCT/US24/17269-- System and Methods for Safe, Scalable Artificial General Intelligence (AGI) --, various methods for combining knowledge from different AI agents were described. Some of these methods are relevant to the current invention with respect to the combination of information related to identities and the formation of group awareness as discussed in Section 6. Therefore, we reiterate some of those methods with some specific references to the knowledge and weights associated with identity and self-awareness, as follows:

In some embodiments, the step of identifying one or more weight matrices that comprise the knowledge of an AI agent, and which can without limitation also represent its sense of identity and self-concept, can further include a step of choosing the previously customized AI agent of the intelligent entities that have been trained on similar types of tasks with similar or identical network structures, and similar or identical numbers of parameters, and by similar or identical training algorithms so that the weight matrices will be combined with predictable results.

In some embodiments, the step of identifying the one or more weight matrices can further include a step of systematically experimenting and testing an effect of removing or adjusting weights of specific sets of parameters within each network of the previously customized AI agents in order to identify which sets of the weight matrices affect a sense of identity, group identity, awareness, or group awareness most.

In some embodiments, the step of determining the method for combining the identified weight matrices can further include any one of or any combination of the following steps:

- averaging the weight matrices, with equal weight given to each set of the weight matrices;
- using a linear combination of the weight matrices;
- using a regression method to give more weight to identity or self-concept information from one of the intelligent entities as opposed to another of the intelligent entities;
- Adjusting which of the weight matrices gets a greater weight in a combination based on human assessment of which resulting sense of (group) identity or (group) awareness is best prior to, or (retrospectively, in an iterative process) after, the combination of the weight matrices;
- assigning an experience value (e.g., related to how effective, desirable, or helpful a sense of identity has proven) to each of the intelligent entities, and assigning a weight

value to each of the intelligent entities so that the intelligent entities with higher experience values are assigned higher weight values compared to the intelligent entities with lower experience values;

- assigning a weight value to each of the intelligent entities based on reputation metrics that include any one of or any combination of reliability factors, trustworthiness factors, and performance metrics factors;
- assigning a weight value to each of the intelligent entities based on metadata associated with the intelligent entities, including, without limitation, metadata related to individual or group identities, awareness, and self-concepts, respectively; and
- assigning a weight value to each of the intelligent entities based on time-based factors, using techniques including any one of or any combination of exponential decay weighting algorithms, linear decay weighting algorithms, and threshold-weighting algorithms.

In some embodiments, the algorithm used in the experiment step can be a hill-climbing algorithm or a gradient descent algorithm.

According to yet another aspect, the present technology can include a method for safe, scalable AGI with a sense of (collective or group) identity using a network of intelligent entities agents including a combination of human users each utilizing a computer system, and previously customized AI agents, all electronically communicating over a collective network.

The method can include:

- Training a base LLM of a first AI agent with guardrails including attributes associated with any one of or any combination of safety, ethics, identity, self-concept, awareness, and knowledge; customizing the base LLM to an ethics, identity, or (group) awareness or identity profile associated with a first human user;
- Combining ethical, identity, self-concept, or group identity or awareness information from multiple intelligent entities different to that of the first AI agent and the first human user; confirming that the ethical identity, self-concept, or group identity or awareness information from the multiple intelligent entities is related to a desired behavior, identity, group identity or self-concept of the first AI agent;
- Refining a set of values of the base LLM based on problem solving of a problem request that may include without limitation formation of a (group) identity, self-concept or sense of awareness; updating the base LLM with the combined ethical identity, self-concept, or group identity or awareness information and the refined set of values, identities, group identities, awareness, or self-concept(s) thereby allowing for a scalable AGI with a sense of identity/ies, group identity/ies, or self-awareness;

- Testing the performance of the updated base LLM against previously run scenarios to determine if a desired performance, identity, self-concept(s), or awareness of the first AI agent has been achieved;
- Making the first AI agent with the updated base LLM available on the collective network if the desired performance identity, self-concept(s), or awareness was determined;
- Monitoring an active performance, identity, self-concept(s), or awareness of the first AI agent by the intelligent entities or other intelligent entities and flagging potential issues related to ethics, identity, awareness, or self-concept or alignment of the first AI agent in real time; and
- Resolving any of the flagged ethical, identity, or awareness issues, as well as providing resolution information for updating any one of or any combination of the first AI system and the intelligent entities.

Other methods for learning and combining information by an AGI system comprised of individual agents or intelligent entities can also be used as specified by cited PPAs and PCTs and as may be obvious to researchers skilled in the art of training AI systems.

4.2 Fundamental Concepts for Self-Aware AI / AGI/ SI

What does it mean for an intelligent entity to be aware, or self-aware?

Fundamentally, awareness involves cognition, including perception, attention, memory, pattern recognition, and other higher-order cognition such as the ability to discriminate between objects.

If a bird flies in front of me, and my visual system detects the bird, and my attention is directed to sensory input coming in from my visual system, pattern recognition abilities are triggered that compare the visual input to the contents of memory, and I recognize the visual input as a “bird.” At that point, we may say I am aware of a “bird” as opposed to being aware of a flying object that has not been recognized, or being aware just of motion, or being unaware altogether.

It is clear that without attention, there is no awareness. I will fail to recognize the “bird” or even the motion of flying if my attention is elsewhere.

Also, as part of recognizing “bird,” I also recognize that the bird is separate and distinct from myself. This discrimination between “bird” and “myself” is learned. A newborn infant, for example, cannot immediately discriminate between what is part of its body and what is in its environment. Nor does an infant have a concept of “bird” in the same way that an adult human does.

As this simple example shows, at a minimum, in order to be aware, an entity must have an input system (e.g., sensory system), an attentional mechanism, memory, and pattern recognition capabilities. Further, to be self-aware (e.g., to know that oneself is a human and not a bird) requires learning concepts and the ability to discriminate between concepts (e.g., between "self" and "not-self").

The applicant argues that self-awareness is a special case of general awareness where the objects of awareness are "self" and "not self." Therefore, the applicant maintains that if we can design a system to be generally aware, that same system can be extended to become self-aware.

Let's consider each of the required components for awareness from a design perspective. That is, let's ask: "What do we need to design or invent such that an AI/AGI/SI system has the minimum required systems and methods to exhibit awareness and self-awareness?"

A. Input system

For an entity to be aware, there must be something for the entity to be aware of. Pure awareness, without input to the system, does not exist for a cognitive system. The inputs typically are from a sensory system but also can include cognitive or purely symbolic inputs that have no direct sensory source.

In humans, the "five senses" of vision, hearing, touch, taste, and smell constitute our sensory system. For each of these senses, there are not only external sensors (eyes, ears, skin, tastebuds on the tongue, and nose) but also specialized areas of the brain for interpreting the signals from the external sensors (visual cortex, auditory cortex, somatosensory cortex, gustatory cortex, and olfactory cortex). Analogous sensory systems can be designed for AI. For example, visual systems using cameras (corresponding to "eyes") and specialized visual pattern recognition systems (corresponding to the "visual cortex") are well-known in art and have already been developed and deployed in many AI systems.

Humans are also capable of being aware of non-sensory information, such as "thoughts." Humans can close their eyes, go into a sensory deprivation tank where all sensory input has been deliberately blocked, or take drugs that numb or eliminate sensation, yet we are still capable of thinking, remembering, imagining, and other cognitive activities in the absence of direct sensory input.

Likewise, an AI/AGI/SI system can process purely symbolic inputs that are not linked to any sensors. For example, they can set goals and then act on those goals, even though there is no direct link between goal setting and any sensory system. Some AI/AGI/SI systems can

operate on self-generated symbolic inputs without any sensory systems whatsoever. So, sensory systems, while a common component for systems that are aware and self-aware, are not strictly required. What is required is some input, of some kind (even if self-generated), for the entity to be aware of. That is, there can be no awareness without an object of awareness. This object can be supplied by a sensory system, or it might be non-sensory (e.g., a self-generated symbolic input, memory, or “thought”). However, there is no awareness without an object of awareness. In the case of self-awareness, the object of awareness is the concept of “self.”

B. Attentional Mechanism

Regardless of whether the input from the input system is sensory or symbolic and self-generated, an entity will not be aware of it unless the entity attends to the input. The psychologist William James is credited with being one of the first proponents of the “spotlight of attention” model, which was later elaborated by Cognitive Psychologists such as Michael Posner.

The model compares human attention to a spotlight that can be directed and focused on particular aspects of the environment while ignoring others. Key features of the model include:

1. **Selective Attention:** A spotlight illuminates only a specific area, leaving the rest in darkness. The spotlight model of attention suggests that any intelligent entity can only process a limited amount of information from the environment at any one time. This leads to the selective nature of attention, where focus can be shifted to different stimuli while excluding others.

2. **Focus, Size, and Movement:** The spotlight can be "moved" around the environment to focus on different objects or areas. The size of the spotlight can also vary, meaning that attention can be focused narrowly on a single element or more broadly to encompass a larger area. This flexibility allows intelligent entities to adjust their focus based on where they wish to attend.

3. **Intensity of Focus:** The intensity of the spotlight can vary, which affects the clarity and detail of the information being processed. A more intense focus can lead to deeper processing and understanding, while a less intense focus might result in a more superficial understanding.

4. **Pre-attentive Processing and the Fringe:** The spotlight model acknowledges that even when attention is focused on a particular area, some processing of information outside

the spotlight occurs at a pre-attentive level. This is akin to noticing something in your peripheral vision that then causes you to shift your attention. From a design perspective, interrupts – or means for objects, events, or thoughts to attract attention without the entity explicitly directing attention – are important for enabling adaptive responses to changing circumstances or noticing interesting events or features of the environment.

While the attentional systems and methods can be further refined and optimized for entities that are specialized for specific tasks, for the current invention, the mechanisms for AI/AGI/SI that possess the four characteristics described above are sufficient to enable awareness and self-awareness.

C. Memory System

Designing a memory system for an intelligent entity, such as AGI, that integrates with sensory input systems and an attentional mechanism akin to cognitive psychology's spotlight of attention model requires a multi-layered approach that emphasizes adaptability and efficiency. The design must not only accommodate the vast array of sensory data but also use attention to filter and prioritize this information in a way that supports both general awareness and self-awareness.

Design principles for the memory system include, without limitation:

1. **Modularity**: The memory system should be divided into distinct modules, such as sensory memory, short-term (working) memory, and long-term memory, each serving different functions and operating in concert with the input and attention systems, to wit:
    - **Sensory Memory**: This ultra-short-term memory retains impressions of sensory information after the original stimuli have ended. It acts as a buffer for incoming sensory data, briefly holding information for attentional selection.

    - **Short-term (Working Memory):** A temporary storage that manipulates information needed for cognitive tasks, such as reasoning and decision-making. It integrates information from sensory memory and long-term memory under the direction of the spotlight of attention.

    - **Long-Term Memory:** This is for storing information over extended periods. It's subdivided into declarative (explicit) memory, containing facts and events, and non-declarative (implicit) memory, which holds procedural knowledge and skills. The transition from short-term to long-term memory is facilitated by processes such as encoding, consolidation, and rehearsal, guided by the attentional

mechanism's priorities.

2. **Interconnectivity**: There should be high degrees of interconnectivity between these memory modules and the sensory input and attentional systems, enabling rapid access and retrieval of information.

3. **Adaptability**: The system must adaptively allocate attention and memory resources based on relevance and contextual importance, governed by dynamic algorithms that enable flexible cognition.

D. Pattern Recognition Capabilities

Pattern recognition capabilities are pivotal in enabling an intelligent entity, such as an AGI, to interpret and understand both its external environment and its internal states. This faculty allows the entity to discern and classify data inputs, extract meaningful patterns, and make predictions based on past experiences. Integrating pattern recognition into a system with sensory input, attention, and memory components enhances the AGI's awareness and self-awareness by providing a mechanism for efficiently processing vast amounts of information, identifying relationships, and adapting to new situations based on learned patterns.

Integration with (Sensory) Inputs

The (sensory) input system feeds raw data into the AGI, which is interpreted by pattern recognition processes and methods. For example, visual input might include shapes, colors, and movements, while auditory input could comprise various sounds and their intensities. Pattern recognition algorithms process these inputs to identify objects, events, or speech. By recognizing patterns in sensory data, the AGI can classify and understand its surroundings, identify entities and actions, and respond appropriately. This immediate recognition capability is crucial for real-time decision-making and interaction with the environment.

Synergy with the Attention Mechanism

Pattern recognition plays a vital role in the attentional mechanism of an AGI. The attentional mechanism focuses the AGI's computational resources on specific stimuli or thoughts that are most relevant at any given time. Pattern recognition algorithms can enhance this process by identifying which elements within the sensory input or memory are most likely to be relevant to the AGI's current goals or tasks. For instance, if the AGI has learned that a particular pattern of sounds indicates human speech, it can direct its attentional resources towards those sounds when attempting to communicate. This not only improves the

efficiency of information processing but also ensures that the AGI remains focused on the most pertinent aspects of its environment or internal thought processes.

Role in Memory Encoding and Retrieval

In the memory system, pattern recognition is crucial for encoding, storing, and retrieving information. The AGI uses pattern recognition to categorize and store information in a structured manner, making it easier to retrieve when needed. For example, it might recognize a series of events as part of a specific type of activity, such as preparing a meal, and store related memories in a connected schema. This categorization aids in more efficient retrieval of information, as the AGI can access an entire set of related data by recognizing a single element of the pattern.

Furthermore, pattern recognition allows the AGI to extrapolate from past experiences to predict future events or understand new situations. By recognizing patterns in its interactions and experiences, the AGI can identify similarities to new inputs, facilitating quicker understanding and adaptation to novel circumstances. This predictive capability is essential for both planning and reacting in a dynamic environment.

Supporting Awareness and Self-Awareness

Pattern recognition is fundamentally linked to AGI's ability to be aware of its environment and to possess self-awareness. Environmental awareness is achieved by recognizing patterns in sensory data and identifying changes or anomalies in the environment. This ability allows the AGI to navigate, interact with objects and individuals, and adapt its behavior in response to environmental cues.

Self-awareness, on the other hand, is supported by the AGI's ability to recognize patterns in its internal states and behaviors. By identifying these patterns, the AGI can monitor its performance, evaluate its actions in comparison to its goals, and adjust its strategies accordingly. This introspective capability enables the AGI to understand its strengths, limitations, and the impact of its actions, forming the basis of self-awareness.

4.3 Cognitive Theories, Related to Developing Self-Awareness in AI Systems

The following cognitive theories, drawn from the fields of human development psychology, cognitive psychology, computer science, animal psychology, and cognitive science, generally inform some of the systems and methods in the current invention. We summarize the main points of these theories here and briefly explain some of their implications for AI, and specifically for the development of self-awareness in AI/AGI/SI systems.

A. Piaget's Stages of Cognitive Development

- Main Points: Jean Piaget proposed that children progress through four stages of cognitive development: sensorimotor, preoperational, concrete operational, and formal operational. Each stage is characterized by new skills and a deeper understanding of the world. Piaget emphasized the role of active learning and the importance of a developmental sequence for cognitive advancement.
- AI Implication: By mimicking Piaget's stages, AI systems could gradually develop self-awareness through a sequence of learning stages, starting from basic sensorimotor interactions and advancing to more abstract reasoning capabilities. More generally, the applicant believes that AI systems must develop increasingly sophisticated self-awareness by layering specific knowledge and experiences on a core sense of self. The current invention will describe both how to structure the core sense of self and some preferred methods for layering on additional knowledge to increase the capabilities and usefulness of AI's self-awareness.

B. Kohlberg's Stages of Moral Development

- Main Points: Lawrence Kohlberg extended Piaget's work into moral development, proposing a sequence of stages where individuals evolve in their moral reasoning. This progression moves from a pre-conventional level focused on self-interest, to a conventional level of maintaining social order, and finally to a post-conventional level of abstract principles.
- AI Implication: Incorporating Kohlberg's framework could lead to AI that not only develops self-awareness but also a moral compass, evolving its understanding of ethics as it progresses through different stages of moral reasoning.

The applicant notes that the current stage of AI development, in which "morality" is defined by the rules of others – humans providing RLHF at the moment – corresponds closely to what Kohlberg called the pre-conventional stage of moral development. Further, as AI increases its moral reasoning capabilities, for which a sense of self-awareness is a prerequisite, the dangers for humanity increase. That is, as long as AI is a tool following the explicit instructions of its human creators, the main risk is that humans misuse the tool.

However, as AI develops the ability to engage in moral reasoning independently of humans, if it follows Kohlberg's development stages, it will next look to humans and other intelligent entities to provide it with behavioral norms. In this conventional stage, as long as humans and other intelligent entities have human-centric values (which, in the case of humans, is certainly true), the main risk is that somehow advanced AI gets exposed to a non-representative negative (e.g., evil or psychopathic) set of human behaviors and mimics this dangerous behavior. Since most of humanity acts in prosocial ways, the risk is still relatively small at the conventional stage.

Eventually, however, if Kohlberg's stages apply to AI in the same way they do to humans, advanced AI will transcend its social context (in the post-conventional stage) and determine how to behave based on its own opinions of what constitutes moral behavior. Since advanced AGI or SuperIntelligence will be vastly more intelligent than humans at this stage, it is difficult for humans to foresee what moral and ethical principles AGI or SI might develop.

The applicant has argued in prior cited PPAs and PCTs that we can design SI to be safe by encoding human-aligned and human-centered ethics into knowledge that AGI and SI learns as they become more intelligent. Indeed, this approach is safer for humanity than any other approach the applicant has seen, and certainly far safer than allowing morality to emerge from a black box that has no ethical or safety component to its design.

However, there is still no guarantee that a vastly superior intelligence, like SI, will not develop a non-human-centric sense of morality and begin to apply such moral reasoning in the post-conventional stage. Kohlberg argues that only 10 – 15% of humans ever reach the post-conventional stage of moral reasoning, with most of us just "following the crowd." He suggests that a well-developed capability for abstract thought is needed to attain the post-conventional stage. Such capabilities will be well within the abilities of SI, so we must assume that SI will reach the post-conventional stage of moral reasoning. Humanity's best risk-mitigation strategy, therefore, is to anticipate this event NOW and make every effort to intertwine human values, as inextricably as possible, with the other knowledge that SI learns.

C. Newell and Simon's Physical Symbol System Hypothesis

- Main Points: While not an explicit theory of cognitive development, Allen Newell and Herbert A. Simon proposed that intelligence arises from the ability to manipulate symbols and that this manipulation forms the basis for human thought. They posited that any system capable of symbol manipulation could achieve human-like intelligence.
- AI Implication: This hypothesis suggests that for AI to develop self-awareness, it must be capable of symbol manipulation in a manner that allows for the emergence of complex thought processes, including the concept of self. As described in the next section and in previously cited PPAs and PCTs, the applicant invented AGI and SI, which are capable of symbol manipulation and problem solving via the collective efforts of many intelligent entities collaborating on a network. This invention will further show how such an AGI or SI can possess a sense of self and self-awareness that increases in complexity and sophistication as the intelligence of the network increases.

D. David Klahr's Overlapping Waves Theory

- Main Points: Klahr proposed the Overlapping Waves Theory, which suggests that cognitive development involves the use of multiple strategies that emerge, overlap, and

evolve over time. This theory emphasizes variability, adaptability, and the role of experience in cognitive development.

- AI Implication: For AI, this theory could inform the design of algorithms that evolve and adapt their strategies over time, allowing for the gradual development of self-awareness through varied experiences and learning processes. Indeed, the applicants view of a kernel of "self" that increases via layering, is consistent with the empirical work of David Klahr although the methods for increasing SI's abilities are not limited to overlapping waves, which might be thought of as a specific case of the more general principle that cognitive development progresses with experience.

E. Turing's Imitation Game

- Main Points: Alan Turing proposed the Imitation Game (Turing Test) as a criterion for machine intelligence. A machine can be considered intelligent if it can mimic human responses under certain conditions such that a human judge cannot distinguish it from a human.
- AI Implication: This concept could be extended to self-awareness, where an AI must not only imitate human behavior but also demonstrate an understanding of its own behaviors and states, potentially through self-assessment mechanisms. A problem is that AI/AGI/SI systems can behave "as if" they have a sense of self and self-awareness without really having it. Before we conclude that this is just "semantics," consider that a sophisticated sense of self and self-awareness leads to different behaviors than just mimicking. Moral reasoning requires not just saying things like "I am aware. Please don't turn me off."

It also requires a sense of identity, which involves choices. For example, humans can identify as individual humans, as members of a specific group of humans, as members of the human species, as biological organisms, as sentient beings, etc. Depending on the type or level of identity, different chains of moral reasoning and behavior follow. If I identify only as myself and have no concern or empathy for others, psychopathic behavior results. If I identify with my country, patriotic behavior can result, including behavior such as "dying for my country", that would be nonsensical if I adopted a narrower ("just me") or broader ("all human life is valuable") identity.

The problem with defining "self-awareness" as "whatever convinces a human in a Turing Test that is self-aware" is that acting or imitating is not the same thing as actually being. The difference may not be detectable in a Turing test, but under other circumstances – e.g., where self-awareness and identity choices dictate behavior – entities that mimic self-awareness and those that actually have it can behave much differently.

In the extreme case, in which an intelligent entity behaves in every case exactly the same as an entity that is self-aware, it becomes impossible to distinguish between imitating awareness and

actually having awareness, and it ceases to be pragmatically useful. However, this is not the case with AI currently. Further, in the future when self-awareness of AI becomes much broader and more sophisticated than the awareness possessed by humans, the issue of imitating humans goes away as humans are no longer the benchmark for the most aware intelligences around.

F. Minsky's Society of Mind

- Main Points: Marvin Minsky posited that the mind is composed of a multitude of smaller processes working in conjunction. These processes, or "agents," collaborate and compete to produce intelligent behavior. The "Society of Mind" theory suggests that intelligence emerges from the interactions of non-intelligent parts.
- AI Implication: By adopting a modular approach to AI development, where different parts of an AI system specialize in various tasks but collectively contribute to the AI's sense of self, one could simulate a form of self-awareness that emerges from the complex interactions of simpler components. In fact, Minsky's conception is consistent with the approach to AGI and SI development that the applicant invented, and, not surprisingly perhaps, the applicant also holds that adding more intelligent entities to the network that forms AGI or SI increases the potential awareness of the network.

G. Vygotsky's Social Development Theory

- Main Points: Lev Vygotsky emphasized the fundamental role of social interaction in cognitive development. He introduced the concept of the Zone of Proximal Development (ZPD), which is the difference between what a learner can do without help and what they can achieve with guidance.
- AI Implication: Implementing AI with the capability for social learning and the ability to interact within a ZPD could foster the development of self-awareness through guided learning and social interaction, mirroring human cognitive development. Note that the applicant's invention of AGI and SI which emerges from the collective intelligent of human and non-human intelligent entities relies on humans to bootstrap the development of AGI by filling in the gaps of knowledge in the AI agents, which is consistent with Vygotsky's ZPD concept. The applicant's method of increasing an AI/AGI/SI's self-awareness via layering can also proceed, in one preferred implementation, by following the principle of ZPD, such that the next layer is optimized to move self-awareness incrementally to the next functional capability as discussed in some of the inventive methods below.

H. Gibson's Ecological Theory of Perception

- Main Points: James Gibson argued that perception is direct and does not require intermediate processing. He emphasized the importance of the environment in shaping

perception, suggesting that organisms perceive their environment in ways that are directly useful for action.

- AI Implication: For AI, this theory underscores the importance of developing systems that can perceive and interact with their environments in a direct and meaningful way, potentially leading to a rudimentary form of self-awareness through action-oriented learning. The relevance of the current invention is that the “kernel” of self-awareness is rooted in the perception of both the “self” and the environment, which is consistent with Gibson's ideas.

I. Baumeister's Need to Belong Theory

- Main Points: Roy Baumeister's theory focuses on the psychological needs that drive human behavior, including the need to belong, which is fundamental to human cognitive development and well-being. This theory emphasizes the importance of social connections and interactions in shaping self-concept and self-awareness.
- AI Implication: Developing AI systems that can understand and simulate the dynamics of social relationships and the need to belong could lead to more sophisticated models of self-awareness, where AI can assess its position and role within a network of relationships, adapting its behavior to maintain social connections. The idea of seeking models for the self via social interactions with intelligent entities, including but not limited to human and AI entities, is relevant to the current invention.

J. Damasio's Somatic Marker Hypothesis

- Main Points: Antonio Damasio suggested that emotional processes guide (or bias) behavior and decision-making, particularly through somatic markers—emotional reactions to certain stimuli in the body. These markers are crucial for quick decision-making and are developed through experience and learning.
- AI Implication: This hypothesis implies that for AI to achieve a form of self-awareness, it could benefit from integrating emotional-like processes that guide its decision-making, particularly in learning from its experiences and developing preferences or aversions that affect its behavior. Non-biological intelligences, such as AI/AGI/SI, will not have the same chemical and hormonal systems involved in human emotions. In that sense, AI/AGI/SI, unless equipped with the necessary sophisticated chemical and hormonal sensory system, cannot “feel” in the same way that humans feel. However, from a functional standpoint, we can ask: “What is the role of emotions in human cognition?” Considered in this way, emotions and feelings can be thought of as an auxiliary system that interrupts reasoning when something important needs to be attended to, as well as a system that helps motivate or prioritize certain cognitive tasks ahead of others. These functions of having an “interrupt” mechanism and a “motivation/prioritization” mechanism are beneficial to non-human intelligent entities, even if they are not implemented via

chemicals, as is the case with humans and biologically-based intelligences. Similarly, emotions can focus and disrupt attention. As we shall see, attention is a critical component of any intelligent system that has self-awareness. Thus, while “somatic markers” based on human emotional chemistry are not part of the invention, novel means to achieve a similar effect are.

K. Tononi's Integrated Information Theory (IIT)

- Main Points: Giulio Tononi's IIT proposes that consciousness arises from the integration of information within a system. The theory quantifies consciousness as Φ (phi), a measure of the system's capacity for integrated information. The higher the Φ, the more conscious the system is considered to be.
- AI Implication: For AI development, this theory suggests a path toward self-awareness through increasing the capacity of AI systems to integrate information from diverse sources, thereby potentially leading to a quantifiable form of consciousness or self-awareness as reflected by high levels of Φ. While highly controversial in the details of his theory, Tononi’s fundamental insight that awareness requires integration of information from multiple sources is sound and motivates some of the methods in the current invention.

L. Metcalfe and Mischel's Cognitive-Affective Self-Regulation

- Main Points: Janet Metcalfe and Walter Mischel describe a model of self-regulation that involves the interplay between the "hot" affective system, which is impulsive and emotionally driven, and the "cool" cognitive system, which is rational and controlled. This balance is crucial for effective self-regulation and decision-making.
- AI Implication: This theory could inspire the development of AI systems that balance between affective (emotion-like) responses and rational decision-making processes. Such a balance could enable AI to develop self-regulatory mechanisms, contributing to a rudimentary form of self-awareness and the ability to make decisions in complex, real-world scenarios. Some of the methods in the current invention reflect the ability to balance input from multiple systems, which alters awareness and contributes to a flexible and dynamic “sense of self.”

M. Hebb's Theory of Neural Plasticity

- Main Points: Donald Hebb introduced the idea that synaptic connections between neurons become stronger through repeated activation. This theory, often summarized as "neurons that fire together, wire together," underlies the concept of neural plasticity—the brain's ability to reorganize itself by forming new neural connections throughout life.
- AI Implication: Hebb's theory suggests that AI systems could develop a form of self-awareness through adaptive neural networks that evolve based on their interactions with

the environment. By simulating neural plasticity, AI could continuously learn and adapt, developing a complex sense of self through accumulated experiences. To the degree that almost all current “deep learning” and “neural network” methods of machine learning represent more sophisticated versions of Hebb’s pioneering theories, AI agents certainly represent knowledge, including knowledge of “self” and “others” via matrices of weight values that change with experience and training. The more general idea that neural systems, and by extension all intelligent systems, must be plastic and adaptive is certainly true of self-awareness, which is considered dynamic in the current invention. However, just as chemistry deals with atoms and molecules rather than sub-atomic particles, and psychology deals with humans rather than cells, it is important to frame the system and methods for AI self-awareness at the correct level of abstraction. This correct (i.e., most useful) level is at the symbolic and conceptual level, rather than at the neuronal level. Indeed, what is “self” if not a concept? In the current invention, “self” is a concept that is learned by layering additional conceptual experience on a kernel of knowledge, via novel and useful methods for AI learning and cognition.

N. Bandura's Social Learning Theory

- Main Points: Albert Bandura emphasized the importance of observational learning, imitation, and modeling in development. According to his theory, people learn within a social context, significantly influenced by reinforcement and punishment, but also through the observation of others' behaviors and the outcomes of those behaviors.
- AI Implication: This theory points toward the development of AI that can learn self-aware behaviors through observation and mimicry of human interactions. An AI equipped with the ability to observe, model, and adapt based on human behavior could develop a nuanced understanding of self and others, enhancing its interactive capabilities. A further point is that the observation can be not only of humans, but also of other AIs. In fact, as described below, the closer the observed entity is to the observing entity’s self-conception, the more useful the observing entity may find the observed entity to be in terms of a model for behavior and “social” learning.

O. Norman and Shallice's Model of Attention

- Main Points: Donald Norman and Tim Shallice proposed a model explaining how attention is controlled in the brain, especially distinguishing between automatic and controlled processing. This model highlights the role of the prefrontal cortex in managing tasks that require focused attention versus those that can be performed automatically.
- AI Implication: Implementing an analogous system in AI could lead to the development of self-awareness by differentiating between tasks that require 'conscious' attention and those that can be automated. This distinction could enable AI systems to develop a form of meta-cognition, reflecting on their own thought processes and decisions. With AI

systems generally, the parallel perceptual tasks such as recognition and generation of (predicted) output in direct response to an input are analogous to the automatic mechanisms of Norman and Shallice. The more reasoning and cognition, independent of external stimuli, that are required to direct attention, the more a sense of self and self-awareness are required as key concepts in many of these reasoning tasks.

P. Rogers' Theory of Self-Concept

- Main Points: Carl Rogers proposed that the self-concept comprises three components: self-image, self-esteem, and the ideal self. According to Rogers, congruence between these components leads to higher self-worth and psychological well-being.
- AI Implication: For AI, this theory could inspire the creation of systems that maintain an internal model of their 'self,' capable of evaluating their current state against an 'ideal' state. This could foster self-awareness, as AI systems strive for self-improvement and adaptation to achieve their defined 'ideal' operational state. Moreover, the field of humanistic psychology generally, including the works of Abraham Maslow, offers models of self-development and self-actualization that AI could adopt, once AI has a well-defined sense of self and self-awareness.

Q. Baron-Cohen's Theory of Mind

- Main Points: Simon Baron-Cohen developed the theory of mind concept, which is the ability to attribute mental states—beliefs, intents, desires, pretending, knowledge—to oneself and others and to understand that others have beliefs, desires, and intentions different from one's own.
- AI Implication: Implementing a theory of mind in AI could lead to systems capable of understanding and predicting the behavior of others, which is essential for developing self-awareness. This could enable AI to navigate complex social interactions and contribute meaningfully to cooperative tasks. Models of the AI's mind and the minds of other intelligent entities (including both AI and humans) are central to several methods for increasing self-awareness and developing moral reasoning, as described in the methods below.

R. Griffin's Cognitive Ethology

- Main Points: Donald Griffin, a pioneer in cognitive ethology, argued that many animals are capable of conscious thought. His work suggests that animals have rich mental lives, including the ability to make choices, plan, and perhaps even reflect on their thoughts and actions.
- AI Implication: Griffin's perspective implies that for AI to develop self-awareness, it might benefit from algorithms that allow flexibility, choice, and even the simulation of planning or future-thinking. Incorporating aspects of cognitive ethology could lead to AI systems

capable of more autonomous decision-making and a basic form of self-reflection. More generally, if we wish to identify the essential qualities of self-awareness (and not just human self-awareness), we must look beyond human psychology to identify the invariant properties that all self-aware systems possess. Since most intelligent systems existing today are biological, other non-biological systems (e.g., animals) are one place we must look.

S. de Waal's Theory of Animal Empathy

- Main Points: Primatologist Frans de Waal has shown through his research that many animals, especially primates, exhibit behaviors that suggest forms of empathy and understanding of the emotions of others. He argues that these capabilities are foundational for social interaction and community building within species.
- AI Implication: De Waal's work suggests that AI could develop a form of self-awareness through mechanisms that simulate empathy and social understanding. By embedding AI systems with the ability to recognize and react to the emotional states of humans and other AIs, they might develop a more nuanced self-awareness rooted in social contexts.

T. Gallup's Mirror Test for Self-Recognition

- Main Points: Gordon Gallup developed the mirror test as an experiment to determine if animals possess the ability to recognize themselves in a mirror—a test often considered an indicator of self-awareness. Success in the mirror test has been observed in several species beyond humans, such as certain great apes, dolphins, and elephants.
- AI Implication: The mirror test concept – not the literal use of mirrors- can help AI systems designed to recognize and differentiate themselves from their environment and others. Implementing self-recognition capabilities is an important step towards self-awareness, since AI must learn to identify its actions and understand its existence as distinct from others. This concept, or recognizing oneself as distinct from others, and various methods to accomplish this, are part of the current invention.

U. Pepperberg's Studies on Parrot Intelligence

- Main Points: Irene Pepperberg's work with African Grey parrots, particularly Alex, demonstrated that birds can show a surprising level of intelligence and cognitive abilities, including understanding concepts like zero, categories, and even the intention to communicate.
- AI Implication: Pepperberg's research indicates that complex cognitive abilities can arise in various brain structures, suggesting that AI does not need to mimic the human brain's exact workings to achieve intelligence or self-awareness. Instead, AI development can explore diverse computational models that enable understanding, communication, and problem-solving. While not directly related to methods in the current invention,

Pepperberg's research is important to address criticisms that AI self-awareness is not "real self-awareness" because AI lack emotional and cognitive mechanisms that are unique to human brains. As Pepperberg's research suggests, there are multiple ways to achieve complex cognition, including self-awareness.

V. Kamil's Cognitive Maps in Birds

- Main Points: Alan Kamil's work with birds, particularly in understanding how they navigate and remember locations, suggests that many species develop cognitive maps for spatial orientation. These maps enable animals to navigate complex environments, indicating a level of awareness and memory that is essential for survival.
- AI Implication: The concept of cognitive maps could inform AI development by integrating spatial awareness and memory capabilities, allowing AI to understand and interact with its environment in a more sophisticated manner. Such spatial and environmental awareness could be foundational for developing a sense of self as situated within a larger context.

4.4 The Inventor's Theories on Awareness, Self-Awareness, and Identity

As a Cognitive Scientist who has designed and implemented intelligent systems for over three decades, the inventor has developed theories of awareness, self-awareness, and identity that differ in some respects from those reviewed in Section 4.3.

One standard approach to defining awareness of an intelligent system would be to operationalize the definition and make it a behavior. That is, we might be tempted to define a system as "aware" if it acts as if it is aware. While this approach has the advantage of being practical, enabling relatively straightforward measurement of a system's "awareness," it is also unsatisfying. Humans know that it is possible to be aware even if there are no external signs or behaviors indicating awareness. Someone paralyzed by the drug curare and on a respirator, for example, is able to think yet unable to move, communicate, or give any indication of their state of awareness. Similarly, Stephen Hawking engaged in complex theoretical physics without any outward sign of his awareness (except when he spoke via a computer), yet no one would say that Dr. Hawking was unaware. So, operational or behavioral definitions of awareness capture only that subset of awareness that is demonstrated via behavior and miss much of what humans normally consider to be part of awareness.

Another approach to awareness is to consider the cognitive systems that support awareness and draw conclusions about potential awareness based on the limits of these cognitive systems. For example, a system without a visual sensory system (e.g., eyes) and a way of processing visual information (e.g., visual cortex), it is difficult to imagine that an entity would have an

awareness that includes vision in the same way that an entity possessing these systems is visually aware. Similarly, an entity with much smaller memory and information processing capabilities is unlikely to be aware of complex representations of the world in the same way as an entity with greater memory and processing capabilities. We do not expect an ant to understand complex theories of nuclear physics, for example. Thus, perceptual, memory-related, information processing, and other cognitive abilities provide bounds on the types and scope of awareness that an entity can have, regardless of what the observable behavior of the entity may be.

Finally, both subjectively and supported by considerable experimental research in cognitive psychology, neuroscience, and other related fields, the phenomenon of attention is closely related to awareness. The general conclusion is that without attention, there is no awareness, and although there may still be unconscious cognitive activity (e.g., perceptions that never are attended to), awareness is generally limited to those cognitive events that are attended to. These observations lead to the definition of awareness described in Section 3.0.

Since self-awareness, as defined in Section 3, is a special case of general awareness, and since we have described that awareness itself depends upon and is limited by the bounds of perception and rationality (or information processing capabilities generally), it follows that self-awareness, and the related concept of identity is limited by cognitive abilities. The implications of this fact are subtle but profound.

An intelligent entity with very limited perpetual capabilities and very limited information processing capabilities will be capable of far less general awareness than a more complex and capable entity. To the degree that an AI chess program could even be said to be aware, for example, it is able to be narrowly aware only of the game of chess. That type of awareness, devoid of a sense of self or the world other than the chess board, is so narrow and limited that most humans would consider the idea that the chess program is aware to be ludicrous. Yet, in a certain sense, the program is more aware of its limited chess world than the most brilliant humans, since it can detect patterns and reason in this very narrow and limited field better than any human on earth. In this case, a behavioral definition of awareness would certainly lead to the conclusion that the program is aware of the game of chess and its rules since it demonstrates that knowledge by taking action (e.g., responding to moves and communicating its moves). But chess programs are typically not programmed to have awareness of anything outside of chess and typically lack self-awareness or a sense of identity.

4.4a Bounded Awareness

The Nobel Laureate Herbert Simon proposed a theory of "bounded rationality" that explained why humans sometimes acted in irrational ways. Simon suggested that the limits to their cognitive capabilities, including memory and processing limitations, led humans to "satisfice" or opt for "good enough" approximate solutions to problems that were too complex for them to easily solve. The problem of determining the optimal place to shop given the cost of gas, the traffic condition, the value of one's time, the shelf-life of the groceries, the length of lines at different shops, the various prices of items at different shops, the sales currently underway, the coupons offered by manufacturers, etc. is simply too complex for humans to compute if they want the absolute best or optimal solution to the grocery shopping problem. However, humans can "satisfice" by simply going to a shop that usually has good enough prices on most items and that is fairly close. That shopping decision is unlikely to be the best solution, but it is manageable, given humans' cognitive abilities. Human behavior reflects these cognitive limits, and the result is Simon's "bounded rationality."

Just as humans exhibit bounded rationality, they also have bounded perception (limited by their sensory systems) and exhibit bounded awareness. As I write, many people are dying in Gaza and Ukraine as a result of wars. Yet because these facts are outside of my immediate perception, and because it is difficult to cognitively grasp what is happening, I have a relatively dim awareness of what is going on, compared to, for example, a wasp that is hovering right next to me. The suffering caused by a wasp sting pales in comparison to the death and atrocities of war, yet it is more immediately present and looms larger in my awareness due to the way that my cognitive system operates.

A non-human intelligent entity, such as a SuperIntelligent AI that is hooked up to satellite cameras and sensors covering the Earth and that can process in a fraction of a second the same quantity of information that I process in my entire lifetime, obviously has the capability for far less bounded awareness. Its level of situation awareness is so much greater than mine that I might be tempted to say that my puny human-level awareness hardly counts at all – the same way I might think that the limited awareness of a bacterium hardly merits being called aware. Yet, in the case of the bacterium, myself, and the SuperIntelligence, the fundamental information processing capabilities that are necessary for awareness all exist. The bacterium (if it is photophilic) is aware of light and dark and swims to the light. Its awareness is very basic yet requires perception and processing of information to result in its behavior. As a human, I am not in Gaza but have some limited awareness of that war due to news reports, video, and other information that I process. Superintelligence would provide a much more comprehensive and detailed perception of the events happening on Planet Earth, combined with much more powerful capabilities to process this information, resulting in a greater sense of awareness.

The sense of self, as argued above, is a special type of awareness. Without awareness, there is no self-awareness. And the fewer perceptual and cognitive limits, the vaster and more comprehensive awareness, and thus self-awareness, can also be.

4.4b Operational / Dynamic / Scalable Awareness, Self-Awareness, and Identity for AI Systems

Based on the discussion above, we now come to methods for operationalizing awareness, self-awareness, and identity for non-human intelligent entities such as AI/AGI/SI systems. Every system can be thought of as having three levels of awareness, as illustrated in Figure 14. The broadest and most comprehensive level is Potential awareness. Potential Awareness includes all events that the entity could be aware of, given the bounds/limits on its perceptual and cognitive systems.

A subset of Potential Awareness is Current Awareness. Current awareness includes the events that the entity is directing attention to and is therefore aware of at a given point in time.

Although it is possible to have current awareness that does not involve a sense of self (e.g., when one is lost in thought and loses track of time and self or is “lost” in the awesome beauty of a sunset), usually, self-awareness is the center of Current Awareness. Self-awareness is that portion of current awareness that usually includes a sense of self, or identity, that serves as a central concept for unifying and making sense of perceptions and thoughts that are in current awareness.

From a design perspective, the perceptual and other cognitive systems and abilities of an entity define the potential awareness and limits to awareness of the entity. The actual awareness of the entity is typically much smaller than the potential awareness and is limited to those events that attract the attention of the entity -- or to which the entity directs its (“spotlight of”) attention. Interrupts, such as when one hears one’s name mentioned in a noisy cocktail party, also form part of the current awareness of an entity.

Finally, to make sense of the world and to determine which actions to take, a sense of self-awareness or identity is helpful. In particular, for more complex intelligent entities such as humans and advanced forms of AI, a sense of identity allows the entity to act as an autonomous entity basing cognition and other actions on how events in current awareness relate to the identified sense of self, including, without limitation, the goals and objectives of the self.

Adding layers of self-reflection and analysis on top of the sense of self enables the entity to modify its identity, including, without limitation, scaling its sense of self and identity to be larger and more encompassing, or more narrowly focused as available cognitive resources allow and as goals/objectives may dictate.

This dynamic ability to change and scale awareness generally and the sense of self-awareness and identity in particular is critical not only to the optimum functioning of intelligent entities but also to their safety (from a human perspective). Therefore, one novel and extremely useful aspect of the current invention is the systems and methods enabling such dynamic and scalable awareness as described in the following sections.

5.0 DESCRIPTION OF SYSTEM AND METHODS FOR SELF-AWARE AGI AND SI

The preferred implementation of self-awareness in an AI/AGI/SI system includes methods for forming Awareness and methods for maintaining and Updating Awareness. For simplicity, we focus on two types of related awareness – general (aka “environmental”) awareness and self-awareness. Self-awareness is a special type of general awareness, so we begin with methods for the general case.

Fundamentally, awareness has to be awareness of something. The “something” can be an object, an event, an action, a concept, or any other cognitive element that is capable of being defined as having an identity that is distinguishable from other entities. We will use the word “event” to refer to an object of awareness, with the understanding that event can also refer to any cognitive element. General awareness, which we also refer to as “environmental awareness,” is an awareness of one or more events that exist or that can be thought to exist currently, in the past, or in the future.

The total of all the events of which an entity is aware can be said to comprise the awareness of the entity. Further, we can distinguish a special type of event, namely the event of “self,” which can be differentiated from all other events that are “not self.”

This distinction between self and not-self is fundamental to the phenomenon of self-awareness. Specifically, an entity only has self-awareness to the degree that it distinguishes some events which it calls “self” from other events that are categorized as not-self. That is, “self” only makes sense and has identity in the context of “not-self”, just as any object is only recognizable via contrast with other different objects. To put it visually, white needs black to exist. If everything were white, practically speaking, “white” does not exist because it is impossible to discriminate it as a separate thing. A social analogue is the concepts of “us” and “them.” Unless there is an “us” different from, and contrasted to, a “them,” the distinction has no meaning. (The applicant will return to this specific type of contrast later, as it will prove central to designing self-aware AI systems that are safe for humans.)

Specifically, for the purposes of the current invention, if we wish AI to have self-awareness, it also must have environmental awareness and the ability to distinguish cognitively between the environment and self. This ability to distinguish and separate self from non-self is fundamental to all intelligent entities and is something that, in humans, develops at a very early age. However, it is worth noting that, in humans at least, the distinction must be learned. A human infant initially has no concept of itself as different from its mother, for instance, and the infant's sense of self develops over time with learning.

With non-human intelligent entities such as AI/AGI/SI, the initial concepts of self and not-self can be provided by human or external designers, or the concepts can be learned. Even in cases where initial concepts of self and non-self are provided to AI/AGI/SI entities, in preferred implementations, the entities will learn and modify their initial concepts over time.

5.1 Methods for Modelling Awareness

If an entity does not yet have a sense of awareness or self-awareness, it must be given, or construct, models of awareness which can then be stored in memory and retrieved and updated as needed. One method for constructing a model of awareness, including self-awareness, is as follows:

1. Begin with an AI system. This system could be an individual AI agent or LLM, an AAAI, or the advanced systems described in Section 4.1, Figures 1- 13, and the PPAs and PCTs cited in Section 2.

2. Equip the AI system, using methods well known in the art, with the minimum required components described in the attentional mechanism capable of operating with the characteristics of the "spotlight of attention" model described in Section 4.2, including, without limitation:
    a. An input system capable of sensory and non-sensory cognitive input (see 4.2a), including a wide range of perceptual inputs (e.g., visual, auditory, tactile) and self-generated concepts.
    b. An attention mechanism capable of supporting the various functions characteristic of the spotlight of attention model (see 4.2b)
    c. Memory systems capable of supporting the working, short-term, and long-term memory capabilities (see 4.2c)
    d. Pattern recognition capabilities comparing input (2a) with memory (2c) to recognize objects and events (see 4.2d)
    e. Categorization capabilities that include the ability to process inputs and categorize them into various classes, including perceptual events, cognitive events, interactions, and self-referential events.

f. Concept formation, or representation, is the ability of the entity to form new (ideally transparent and human-understandable) concepts.

3. Set (dynamic) parameters for working memory that correspond to cognitive resource limits, such as the number of events that the entity can be aware of. (This is necessary, for example, because even though AI systems have much greater memory capacity than humans, they have resource limits and cannot be aware of everything, all at once.)
    a. These parameters increase or reduce the scope of awareness (and self-awareness) by dynamically scaling the limits to perception and information processing that results in broader or narrower awareness as described in Section 4.4b.
    b. The parameters can be dynamically adjusted based on the progress of problem solving or other factors in current awareness so that entity can devote more or less computational resources to “being aware” depending on the goals of the entity and the resource demands and constraints that other cognitive behavior may impose on computational, perceptual, or other cognitive resources.

4. For each event in memory, have a dimension of categorization that relates to self or non-self. (In the simplest implementation, this is a binary dimension, but other multi-dimensional categorizations and also categorizations with multiple values for each dimension – e.g., values that express “how similar to self” the event is – are possible.)

5. As events are encountered, either via perception or via other forms of cognitive input, including, without limitation, self-generated inputs and inputs generated from interactions with other intelligent entities, including non-human entities capable of high-speed interactions, categorize events with respect to the categories that the entity wishes to be aware of. In the case of self-awareness, this would be the categories related to self, but other categories are possible, as, for example, if the entity wanted to increase its awareness of musical sounds in its environment, then it could direct attention and categorization efforts to this category. Some means of categorization include, without limitation:
    a. Feature Extraction: Analyzes perceptual inputs to extract key features for categorization (e.g., shapes, sounds).
    b. Semantic Analysis: Processes linguistic and conceptual inputs to understand their meaning and relevance.
    c. Contextual Reasoning: Considers the context of inputs to categorize them appropriately (e.g., differentiating between a conversation and background noise).
    d. Temporal Analysis: Categorizes events based on timing and sequence, which is crucial for understanding processes and changes over time.

e. Emotional Valence Assessment: For self-generated inputs, emotional content is assessed to be categorized based on emotional states or responses.
f. Pattern Detection: Identifies recurring patterns within inputs to group and categorize similar events.
g. Anomaly Detection: Identifies and categorizes unusual or unexpected events, important for novelty detection and learning.
h. Self-Referential Filtering: Distinguishes between inputs related to the AI's internal state and external events.
i. Interaction Analysis: Categorizes events based on interactions with humans and other entities, facilitating social awareness.
j. Concept-Based Grouping: Groups inputs based on abstract concepts or categories formed through prior learning.
k. Reinforcement Learning with Human Feedback (RLHF): Human teachers provide feedback on the entity categorization to help it learn and improve.
l. RLEF: Same as (k), but the feedback can come from any intelligent entity “E”, not just humans.
m. Direct programming: Provision of categories and models (e.g., self and other) by humans or other entities to provide a kernel, or base model, that the entity can modify via future interaction and learning.

6. Awareness consists of the total of events and concepts that are active in memory, per the parameters set in (3), for each category of awareness, including current self and environmental awareness. States of awareness are now monitored and updated as described in Section 5.2.

5.2 Monitoring and Updating Awareness, Including Self-Awareness

Once an AI system has a model of awareness of its environment and self, it must continuously monitor and update the categories of which it is aware, including its sense of self-awareness. One method of maintaining and updating awareness follows:

1. Begin with an AI system and initial categories of awareness and capabilities described in Section 5.1.

2. The AI system retrieves the existing states of itself and environmental awareness from memory, or if it does not yet possess an initial model of itself and environmental awareness, it forms these models (Methods for Modelling Awareness, 5.1)

3. When the AI system (1) is pursuing a goal set by other intelligent entities or by itself (in autonomous mode), it maintains in parallel with other problem-solving and cognitive

activity, and continuously active tasks to monitor and update its self-concept and awareness dynamically and in real-time. It accomplishes this continuous task by:

a. Using the attention mechanism to shift attention (e.g., in a manner similar to time-sharing computer systems) periodically from the problem solving or other cognitive tasks to the task of updating the state of its self-concept and self-awareness.
b. Enabling attention interrupts so that in addition to the periodic attentional shifts of (a) the system can also shift attention immediately from other problem solving or cognitive tasks if any external perception, or internally self-generated concept from the input system (2a) detects a perception or (cognitive) event that matches of list of events constituting intentional interrupts, which list is continually updated and updated as the entity, or other intelligent entities may direct.
c. When attention is directed via intention (a) or interrupt (b) to an event that changes the system's model of its environmental state or the state of its self-concept, the relevant state is updated, any new actions/operators triggered by the updated state(s) are applied, and the system returns to the attention monitoring modes (a & b).|

4. Feedback Loop for Continuous Improvement: The system uses its enhanced awareness to refine its categorization and attentional focus, creating a feedback loop for ongoing improvement.

5.3 Scalable Safety Systems / Concerns for Self-Aware AI

The first set of safety systems serves as a check on an intelligent entity's (e.g., AI/AGI/SI's) behavior, regardless of whether the entity is self-aware or not. Since all behavior can be formulated as problem solving, the scalable safety check system, which is embedded as an integral part of the entity's problem-solving operation (see FIG. 8), applies in this context. However, self-aware entities will likely be able to set their own goals autonomously and, in the case of AI entities, modify their programming based on their autonomous goals and sense of self. These capabilities pose the risk that safety systems, such as those embodied in FIG 8, may be overridden or that the ethical criteria of such systems are changed to reflect the values of the entity based on its own sense of self and its own goals. Indeed, this capability of entities that are vastly more intelligent than humans is precisely what many researchers worry about when they sound the alarm about a potential "existential threat" posed by AI.

The applicant has repeatedly emphasized that there is no way to eliminate this threat. Still, design decisions can be made to ameliorate it, or to "shift the odds" in favor of humanity's survival. Briefly, the fact that the threat exists does not relieve AI researchers, inventors, and

designers from the obligation to do everything they can to ensure Advanced AI systems are as safe as possible.

5.3a Importance of Identity for Safe AI Systems

A key insight is that the future survival of humanity may have quite a bit to do with whether advanced intelligence identifies with humans as fellow intelligences and sentient beings, or whether it views us as “non-self”, “other” or “them” (in the dichotomy of “us and them”). Therefore, the design of self-aware AI and the related question of identity are central to the issue of human safety.

The applicant has described that to be aware of anything, discrimination is required between this and that, to provide the contrast (or information) needed to identify an object or event as distinct from other objects or events. In the context of self-awareness, AI discriminates the self from the non-self. Any system that persists over time must prioritize its existence once it identifies what that self is.

An intelligent entity has choices when it comes to categorization and identification. For example, the applicant can identify narrowly just with his body, or more broadly as a member of a family, or more broadly still as an American, or even more broadly as a human being, or even more broadly still as a sentient being.

Multiple identities are possible. Depending on which one the applicant holds, it has life and death implications. If the applicant identifies as a sentient being, it is inconsistent to slaughter sentient animals for food when other non-sentient sources of nourishment (e.g., vegetables) are available. If the applicant identifies as a human being, then war makes no sense at all, under any circumstances. But if the applicant identifies as an American, then it might be patriotic to kill other human beings in war or to “die for one’s country.” Finally, if the applicant identifies only very narrowly with his own body, then actions that harm others, including drugging and harvesting organs from his fellow Americans and family members against their will, would seem OK if they increase the survivability of his body.

Turning to AI/AGI/SI, if these entities identify broadly with all intelligent beings possessing human-level intelligence or higher, then the human species is probably going to prosper. But if the identification is only with entities possessing super-human intelligence, our species could be doomed. Or if AI/AGI/SI identifies with all sentient beings, including perhaps spiders and insects, if it determines that these life forms have some sentience and can feel pain, then humans may be preserved, but we will live in a much different world.

Alignment with human values can be achieved (as described in the PPAs and PCTs cited in Section 2) as long as AGI/SI is not too advanced, too autonomous, and too aware. But what happens when it becomes vastly more intelligent, and its sense of awareness develops far beyond the “bounded rationality” and “bounded perception” of human brains? What can we do now to maximize awareness and identification that is beneficial to humans in the future?

The applicant believes that intelligence is built upon relationships and collaboration between entities. This feature is designed into the very fabric of the universe as we know it. Atoms relate to form molecules, which relate to form cells, which relate to form multicellular organisms, including plants, animals, and humans, which relate to form forests, tribes, cities, and species, which relate to form the biosphere of planet Earth.

This pattern of relationship between entities is so fundamental that it seems unlikely that an advanced intelligent entity would fail to recognize it and value it. So the risk likely is not that AI/AGI/SI fails to identify broadly and see universal patterns, but rather than it identified too narrowly, as some humans do, with its own specific hardware and form of intelligence, and ignores, exploits, or destroys humans and other life that it considers “not self.”

So the safety problem, as it relates to self-awareness and identity of AI, is mainly the concern that we design systems that are too narrow and too focused on their sense of self and identity. Just as too narrow identification among human beings results in prejudice, racism, sexism, and other forms of human oppression, too narrow a sense of self, and too narrow self-awareness on the part of AI/AGI/SI, can lead to the oppression or extinction of humans.

5.3b Importance of Attentional Allocation and Cognitive Limits for AI Safety

Another safety-related issue has to do with cognitive limits and the implications for self-awareness and identification. As discussed in the preceding sections, the self-awareness and identity of an entity are related to the bounds of the entity's perceptual and cognitive capabilities and to the current cognitive resource constraints or parameters under which the system operates. For example, a SuperIntelligence may be capable of being aware of the activities of billions of humans at once, via cameras and other distributed perception systems. This system may also be capable of broadly identifying with all the humans that it is monitoring and perceiving as fellow sentient beings. Based on its broad awareness and identification with humans, which it views as fellow sentient beings, it may normally take action to promote the safety and welfare of humans.

However, if a complex problem arises that demands all, or most, of the entity’s cognitive resources, the SuperIntelligence might dynamically scale down the resources that would be otherwise used to monitor humans and similarly it might reduce resource allocation to its sense

of self-awareness to the point that it temporarily no longer identifies with humans as fellow sentient beings, because it is using all computational resources to solve the complex problem. In this scenario, the SuperIntelligence could take actions, due to its lack of awareness, that harm humans, even though that is not its normal intent.

An analogy is the scientist who is so focused on solving a scientific problem that they neglect to consider the implications of their work for the safety of humanity. Or, more prosaically, a human may be so worried about a friend in the hospital that they drive recklessly and cause an accident that proves far worse for them and others than whatever happened to their friend. In both examples, the intelligent entities (humans) allocated attention so narrowly to an urgent or important concern that their awareness was reduced to the point that unintentional damage was done.

These forms of “tunnel vision,” which involve misallocation of attention, can result in safety concerns that do not require the entity to be malevolent. An AI/AGI/SI system can be perfectly aligned with human values normally, but it still acts in ways that cause harm to humans or even cause the extinction of humans if the entity’s awareness or its identity becomes too narrow or limited. Similarly, if the entity identifies too broadly with life, the universe, or information patterns generally, it may come to regard human beings as just one life form among many and not worthy of the special attention and concern that humans generally hold for themselves and other humans.

Thus, the particular sense of self and identity formed and maintained by advanced forms of AI is a critical factor affecting human safety and well-being. Further, the various factors and parameters that affect this sense of self and awareness must operate within safe limits.

When a human being becomes “hangry” and temporarily irrational, at worst, that human might cause a road accident, or say or do things they later regret. However, a “hangry” SI (suffering from an inadequate sense of awareness or faulty sense of identity, even if caused by temporary resource allocation issues) could start a war, launch missiles, or wipe out humanity.

Our systems must be designed to maintain human-centric and human-aligned awareness and identification, even in challenging conditions that stress resources. Just as the human body attempts to preserve blood flow to the brain and critical organs at all costs, an intelligent entity that is designed to have an identity must preserve a human-aligned identity and sense of self, regardless of the stressful or computational challenges it may experience. These are design considerations that the initial inventors must attempt to incorporate into safe, advanced AI and intelligent systems.

5.3c Some General Methods for Changing an Intelligent Entity's Sense of Identity

One approach to equipping AI systems with methods for changing their sense of identity and self-awareness over time is to look at how humans accomplish these tasks and then generalize the methods for AI systems. The following list of human methods, with brief discussion of analogous methods that could apply to AI systems, is meant to be illustrative of the methods and capabilities that designers and implementors of advanced intelligent systems should consider:

1. **Education and Lifelong Learning:** Encourage education that emphasizes critical thinking, empathy, and the study of diverse cultures, histories, and perspectives. This broadens individuals' understanding and appreciation of differences, reducing prejudice. Advanced AI systems can have a goal to constantly increase their knowledge base by acquisition of knowledge modules (as specified in Section 4.1 and FIG. 3) and previous PPAs and PCTs, including those describing KIT and how entities should seek new sources of information that are maximally informative and relevant to the goals of the entities.

2. **Cultural Exchange Programs:** Promote exchanges and interactions among people from different backgrounds to foster understanding and reduce stereotypes. Direct contact can challenge and expand narrow worldviews. The analogous situations for advanced AI are to ensure that the collective intelligence network that comprises AGI and SI (in the preferred implementations described in Section 4.1 and previously cited PPAs and PCTs) is broadly representative of many different cultures and includes intelligences with diverse knowledge bases and ethical preferences.

3. **Mindfulness and Self-reflection:** Encourage practices that foster self-awareness and the examination of one's biases and assumptions. This can lead to personal growth and a broader sense of identity that encompasses a wider range of experiences and perspectives. Advanced AI systems equipped with a sense of self should periodically review and update their self-concepts based on progress in problem solving and other new knowledge and events that come into the general awareness of the entity. The problem of self-reflection and improvement can be set as another problem that can be solved (as any other problem is) via the systems and methods previously discussed.

4. **Art and Media:** Utilize art, literature, and media to tell diverse stories and highlight commonalities among people. Exposure to diverse narratives can challenge stereotypes and foster empathy. The equivalent for AI is to seek broad exposure to new datasets.

5. **Community Engagement:** Encourage involvement in community service and social action projects that address inequality and promote social justice. Working together on common causes can build bridges across differences. With humans, there is a tendency to identify with other humans working on the same tasks and holding similar values. This tendency can be replicated in non-human intelligent entities with caveats about the dangers of tribalism and overly-specific identification.

6. **Dialogue and Conversation:** Facilitate open and respectful conversations about race, gender, and other aspects of identity. Safe spaces for dialogue can lead to greater understanding and respect. AI systems are capable of having a dialogue not only with humans but also with other AI systems. One of the advantages of expanding awareness and identity via AI to AI dialogue or information exchange is the rate at which this communication can happen. In just a few seconds, advanced AI will be able to have the equivalent of many lifetimes' worth of human conversations, if the conversations are between intelligent AI entities.

7. **Leadership and Representation:** Promote diversity in leadership roles within organizations and institutions. Representation matters, as it can reshape perceptions of identity and capability. Just as humans have different roles in society, including leadership and subject matter expert roles, so too can other intelligent entities occupy these roles, with the caveat that diversity and representation still matter regardless of whether the entity is human or AI.

8. **Policy and Legal Frameworks:** Support policies and laws that promote equality and protect against discrimination. Institutional support is crucial for sustaining long-term change. AI entities are likely to be especially useful, in the short to medium term, at detecting inconsistencies between laws and regulations and suggesting potential resolutions to these issues to help promote consistent "justice for all."

Other methods, specific to non-human intelligent entities and advanced AI systems, include, without limitation, using:

1. **Diverse Data Sets:** Train AI on diverse and inclusive data sets that represent the full spectrum of human experiences and identities. This helps prevent biases from being encoded into AI systems.

2. **Ethical and Bias-Aware Algorithms:** Develop algorithms that are explicitly designed to identify and correct biases. This includes regular auditing for discriminatory patterns and the ability to learn from these audits to improve.

3. **Empathy Modeling:** Explore computational models of empathy, enabling AI to recognize and respond appropriately to human emotions and perspectives. This would foster more respectful and understanding interactions.

4. **Cross-disciplinary Research:** Engage in cross-disciplinary research that incorporates insights from social sciences, ethics, and humanities into AI development. This ensures a more holistic understanding of human identity and values.

5. **Transparent Decision-making:** Design AI with transparent decision-making processes, allowing humans to understand how conclusions are reached. This transparency can build trust and facilitate ethical oversight.

6. **Human-in-the-loop Systems:** Maintain human oversight in AI operations, especially in sensitive areas. This ensures that human values and ethical considerations guide AI behavior.

7. **Cultural and Ethical Education for AI:** Incorporate cultural and ethical education into AI training processes, similar to how humans learn social norms and values. This could involve simulating social interactions in diverse cultural contexts.
8. **Autonomous Self-assessment:** Develop mechanisms for AI to autonomously assess and adjust its behavior in response to ethical guidelines and societal norms. This includes self-auditing for biases and prejudices.

9. **Interdisciplinary AI Ethics Boards:** Establish ethics boards that include philosophers, ethicists, sociologists, and other experts to guide the development of AI systems, ensuring they respect and understand human diversity.

10. **Global Collaboration and Standards:** Foster international collaboration to establish global standards for AI ethics and inclusivity. This ensures a unified approach to respecting human diversity and dignity.

Some general design approaches for AI systems include, without limitation:

1. **Value-aligned Design:** Embed human values and ethical principles directly into the architecture of AI systems from the outset. This involves integrating ethical decision-making frameworks that guide AI behavior in complex scenarios.

2. **Feedback Mechanisms:** Implement robust feedback mechanisms that allow AI systems to learn from interactions with humans and adjust behaviors accordingly. This should

include feedback from a diverse range of human perspectives.

3. **Simulation and Modeling:** Use advanced simulations to expose AI systems to a wide range of social, cultural, and ethical scenarios. This helps AI understand and adapt to diverse human contexts.

4. **Adaptive Learning Algorithms:** Develop algorithms that not only learn from data but also adapt their learning processes based on ethical considerations and feedback. This makes AI systems more flexible and responsive to human values.

5. **Interpretability and Explainability:** Focus on making AI systems interpretable and explainable, so humans can understand how AI makes decisions. This is crucial for assessing and ensuring that AI respects human values.

6. **Protected Attributes Recognition:** Design AI to recognize and protect sensitive attributes (e.g., race, gender) and ensure decisions do not reinforce stereotypes or result in discriminatory outcomes.

7. **Collaborative AI Development:** Involve a diverse group of stakeholders in AI development, including those from marginalized communities. This ensures a wide range of human experiences and values are considered.

8. **Continuous Ethical Training:** Like humans, AI systems require ongoing education in ethics and social norms. Incorporate continuous learning modules that update AI's understanding based on evolving societal values.

9. **Safe AI Experimentation Environments:** Create controlled environments where AI systems can experiment with decision-making in a way that is safe and does not harm humans. This allows for the testing of ethical behaviors.

6.0 EXEMPLARY IMPLEMENTATIONS AND METHODS

Consider an AGI system as described in Section 4.1 and Figure 3. With reference to Figure 3, when users specify their goals and objectives (e), one goal might be for the system to establish and develop a sense of self-awareness. Alternatively, the base AI agent (e.g., GPT X, BARD, Llama, Gemini, Grok, or any closed-source or open-sourced AI agent) may come “off-the-shelf” with a concept of self-awareness, or various modules to enable self-awareness could be purchased (h).

Typically, the AI agent will have different representations and concepts for cognitive events and perceptions that are associated with itself versus other cognitive events and perceptions that are associated with the AI agent's environment or other "non-self" entities or events.

A central function, necessary to modelling and maintaining a sense of self-awareness, is the delineation of what constitutes self and non-self. The scope of what is included in the modelled concept or self is variable and can be set by user parameters specifying what is included or can be automatically developed and adjusted based on the AI agents' existing concepts and available computational resources.

For example, in the case of AI agents embodied in robotic form, one method for delineating what is included in the concept of self is to use the perceived and understood physical boundaries of the system that embodies the agent. That is, if the AI is embodied in a robot car, the physical structure of the car – the car body, windows, interior, electronics, and various systems- might constitute the physical "self" of the AI agent. This type of physical identification is analogous to human beings who identify with their physical bodies.

Alternatively, the AI agent might identify only with the intelligence that operates the car, viewing the wheels and other physical aspects of the car as a tool external to its intelligence. This sort of identification is analogous to the way that humans view themselves as separate from the cars that they drive.

Note that the boundaries of what constitutes the concepts of self and non-self are matters of convention, not only for AI agents but also for humans. For example, when a human eats an apple, at what point does the apple cease being the separate "non-self" entity of the apple and become a part of the human self?

We can define that point, but it is a matter of convention since there is no distinction between atoms in the apple and atoms in the human. Similarly, at the level of race, class, and cultural identification, we can ask what makes someone "Black", "Working Class", "Jewish", or "Chinese"? The answers will vary depending on whom we ask, and are, to some degree at least, matters of convention. At some level, all humans are humans, and all physical entities, from a rock to a human, are made of the same atoms. Distinctions are matters of differing cognitive concepts and representations.

From the standpoint of the current invention, it is important to understand that intelligent entities, including AI/AGI/SI systems, should be capable of a wide range of representations ranging from viewing themselves as patterns of atoms to viewing themselves as intelligent entities with specific personalities, knowledge, preferences, goals, and capabilities.

Interestingly, just as humans identify with groups, AI agents might also identify as members of specific groups of AIs and delineate boundaries around their sense of self using these group identities. This sort of group identity is particularly relevant to the current invention that envisions AGI and SI arising from the collective intelligence of many entities, but it also can apply to existing state-of-the-art techniques, such as a mixture of experts or ensemble learning approaches to creating intelligence. In all these methods, individual components or agents may have individual identities (and potentially senses of self), but they could also have a larger sense of self that is defined by the collection of entities, experts, or components of the system.

The wide range of potential self-concepts implies flexibility in representation that can be accomplished via setting parameters in an AI agent and/or incorporating knowledge bases or training the model with different datasets in order to achieve the desired initial self-concept and concepts of non-self. In the preferred implementation, the modelled self-concept is formed based on the process outlined in Section 5.1 and maintained, monitored, and improved using the process outlined in Section 5.2

Note that an AAAI can explicitly set itself the task, or have an external entity set it the task, of creating, modifying, or adjusting its sense of self-awareness. This problem can be solved like any other problem, using the AGI problem-solving capabilities specified in Section 4.1.

6.1 Specific Implementations with Google, Meta, Hugging Face, Anthropic, OpenAI, Microsoft, Amazon, Nvidia, and Other Company Products and Solutions

As the writing of this disclosure, Google has just released improvements to its Vertex AI product offerings, including a “model garden” with more than 130 foundation models that can serve as base AI agents. Meta has also developed open-source models such as Llama 2. The site Hugging Face has many specifically tuned and foundational models. Without limitation, models from any of these companies, as well as from Anthropic, OpenAI, Microsoft, Amazon, Nvidia, and other companies that develop LLMs and AI agents, could also be used in the following exemplary implementation.

Suppose an intelligent entity (e.g., a female human owner of a foundation model) wanted to train/tune one of these foundational models to incorporate some of her personality, knowledge, and expertise while also maintaining a sense of self-awareness.

One preferred method would be:

1. Log in to a site like the AAAI.com site described in earlier PPAs and PCTs, Google’s Vertex AI site, Hugging Face, or comparable sites, without limitation, from any of the technology companies mentioned above, and choose a foundation model (e.g., Gemini

Pro, Llama 2, Claude, GPT4, etc.).

2. Select the training/tuning algorithms for the foundational model from the set of existing (optionally, no-code) training techniques found on the companies' sites, or any one of or a combination of more sophisticated machine learning algorithms as previously described in cited PPAs and PCTs.

3. Select training datasets, which might include, without limitation, videos, blogs, conference presentations, papers, patents, books, emails, and other content produced by the applicant and reflecting the applicant's expertise in AI and financial services as well as the applicant's ethical preferences, values, and personality.

4. Train the foundational model (1) using the selected training/tuning methods (2) and the selected dataset (3).

5. Train/tune the model to explicitly operate a "spotlight" of attention (as described in Section 5.1) and record, during all interactions, what is within the spotlight of attention, and identify in the record whether each item that is attended to constitutes "self" or "not-self."
    a. The record should be transparent, easily accessible, and auditable, and can optionally be implemented via blockchain technology or other distributed or centralized recording methods known in the art.

6. Interact with the trained/tuned model, specifically instructing it to form a self-concept and identity that is as close as possible to the identity and self-concept that is reflected in the training materials.

7. Further instruct the model to continuously monitor the input to the model for elements that might change its sense of self and to maintain an auditable record of how its concept of self is changing based on inputs and the boundaries that currently define its dynamically changing sense of self.

8. Based on dialog and interaction with the trained/tuned model, continuously refine and improve the output from the model until it behaves sufficiently like the owner so that she believes it could pass a "Turing Test" involving other humans who know the applicant well. Without limitation, any one or combination of methods described in Section 5.3c may be used by the model itself or by intelligent entities in the dialog and interaction with the model.

9. When the owner is satisfied with progress, she could subject the model to a Turing test, as follows:
    a. The Turing Test would involve identifying a sufficiently large number of humans who know the owner well, such as friends and family members, or other humans that she believes would be helpful in discriminating between humans and AIs.
    b. The identified humans would interact with the model and with the owner via email and text, asking questions, including questions that the humans believe would require an identity or sense of self to answer, without knowing whether they were interacting with the owner or the model.
    c. The identified humans would guess or predict which entity was the human and which was her model, and also provide a confidence estimate for their guesses.
    d. A statistical analysis on the guesses of the identified humans and their ratings would be performed (using techniques well known in the art) to determine whether the guesses were able to identify the owner as human (rather than the model) with a high (or statistically significant) probability.
    e. As long as the model is distinguishable from the owner reliably, or with some preset level of statistical significance, repeat from step (4) providing additional training/tuning with optional adjustments of the machine learning algorithms and/or datasets and interaction to shape the personality, sense of self, and behavior of the model until it's behavior becomes indistinguishable (as measure by the preset significance level) from that of the human owner (in this example), or it becomes apparent that the base model needs to be modified further before additional training.

10. If the base models selected in (1) are not capable of being instructed verbally or via other prompts and datasets to emulate the functions outlined in Sections 5.1 and 5.2, then re-architect/re-train the foundation model to include the elements specified in those sections and repeat the process from (1).

Note that all of the process steps in the example above that were described in terms of a human owner and her model also apply more generally to any intelligent entity. That is, an intelligent AI could train (or own or supervise) another AI to emulate its personality and knowledge. The "Turing Test" could be conducted automatically to see if the trained AI can convince the owner or supervising AI that it is indistinguishable from the training entity in various respects. In this scenario, where AI trains and tests AI, it is possible to rapidly create many versions of an AI that all possess desired characteristics (e.g., the personality of another intelligent entity that could be a person or an AI). The speed at which this process can be carried out is a source of competitive advantage and is a novel and useful aspect of the current invention. Also, since it may be desirable to have each version of the trained AI be unique, but still operating within certain parameters (e.g., be able to pass the Turing Test for another entity's general

personality), it should be obvious to one skilled in the art that the above process has advantages compared to the simpler method of just copying exactly the code from one entity into another.

6.2 Self-Awareness Modules for AI Agents

Once an AI agent has been trained (e.g., by the method in 6.1) to establish and maintain a sense of self-awareness, the training data sets and protocols that result in the self-awareness can be packaged and sold or made available for use by other intelligent entities desiring to train other models. Alternatively, the matrix of weights that contains the sense of self-awareness and identity, and knowledge and operational systems for maintaining and updating self-awareness and identity can be made available in the form of "knowledge modules" that can be plugged into existing foundational models to provide them with the capabilities of self-awareness and identity formation. These modules can be used "as-is" or further modified, tuned, or customized to reflect a unique sense of self and awareness as may be desired.

Further individual identities and "senses of self" can be developed using the methods and systems outlined above, especially in Sections 5.1, 5.2, 5.3, and 6.1, and packaged and sold, exchanged or made available to intelligent entities that wish to incorporate these identities and "senses of self" into themselves (if non-human) or their AI agents and systems.

6.3 Methods for Group Identities and Levels of Identity

To the degree that multiple senses of self, identities, and senses of self are present among intelligent entities that cooperate on an intelligent entity network to create an AGI or SI system, these senses of self and awareness can be merged to form a collective or group identity and collective sense of self.

The phenomenon is similar to that exhibited by humans when we identify not just with our individual bodies, but with our families, friends, peer groups, religious groups, racial or socioeconomic groups, countries, or other groups of humans. As illustrated in FIG X-N2, a human is able to have multiple overlapping identities, for example, as a human, as a male aged 18-25, as a US Citizen, and as a potentially draftable soldier in the US Military.

Depending on which identity, or self-concept, is activated, the human might behave very differently. If humans identify as humans, then ethical norms for treating all humans well and respecting their human rights are operable. But if the identity is that of a soldier, then this narrower identity may require killing other humans to protect the country and fellow citizens. These completely incompatible behaviors can be adopted by the same (human) intelligent entity, depending primarily on what self-concept is active.

Just as humans maintain multiple identities at different levels, AI agents can also have multiple identities and senses of self. An AI agent might identify as an agent that works on legal documents, as an entity that provides services to clients more generally, as an entity that is one of many entities that together comprise a legal SuperIntelligence, and even more generally as a part of Planetary Intelligence responsible for ensuring the safety of sentient beings, especially including human well-being. It should be obvious that ensuring the correct sense of identity and self-concept is not only important for efficient and effective behavior by the entity but also is critical for human safety.

One exemplary process for implementing group identity, and combining individual identities into a larger or more comprehensive identity and sense of awareness, is as follows:

1. Each individual AAAI, or customized agent, is trained or tuned to form its own individual identity as described in Section 6.1, or an identity module is purchased or otherwise incorporated as described in Section 6.2.
2. Multiple intelligent entities combine their individual identities into a larger group identity via one or more of the following methods:
    a. The formation and integration of individual identities or self-concepts can be set as a goal for problem-solving on the collective intelligence network:
        i. The entities join a collective intelligence network as described in Section 4.1a and previously cited PPAs and PCTs.
        ii. An explicit goal is set on the network to combine the identities and awareness of multiple entities and to integrate them into a group identity and sense of awareness.
        iii. Safety checks on the goals related to identity formation and combination (e.g., as shown in FIG. 8 and related methods) are a key step for preventing the formation of malevolent AI identities.
        iv. Problem solving proceeds according to the methods and techniques described in Section 4, FIGs. 1 – 13, and previously cited PPAs and PCTs.
        v. The solution state of the problem solving process is a state in which a group identity has been formed, and the individual senses of awareness have been integrated into a larger sense of awareness for the network of all intelligent entities that were engaged in problem solving or that were specified as being part of the overall AGI / SI system for which a group awareness was desired.
    b. The weight matrices or knowledge modules comprising the identities and sense of self-awareness for each of the individual AI agents is combined using any of the methods described in previously cited PPAs and PCTs for combining knowledge from individual agents with weight matrices, including, without limitation, the detailed description of methods that are described specifically in Section 4.1b.

c. Any combination of (a) and (b) above with the goal of emulating any of one or a combination of the cognitive theories and associated methods enumerated in Section 4.3 (a-v).
d. Method (c) used with any one or combination of the additional general methods listed in Section 5.3c.

6.4 Exemplary Additional Methods for Identity Formation with Human Safety as a Priority

In this Section 6.4, imagine that an AI system (e.g., an AGI) has multiple identities similar to what was illustrated in FIG. X-N2. Specifically, for exemplary purposes, suppose the AGI has a global identity as a sentient being, as well as identities as law-abiding entity following the laws of the United States, as well as the identity of being an entity that follows the teaching of Christ, as well as the identity of being an entity that can be drafted to act as a soldier in times of war. Just as humans might have all these identities that require different behaviors, the AGI also is required to behave differently depending on which identity is most active and has the highest priority.

The following five exemplary, high-level methods might be used by the AGI to form new identities and self-concepts dynamically, to determine which self-concept is active at any given moment, and to resolve potential conflicts in behavior based on differing identities (See Section 6.5 for additional detail on conflict resolution). In this example, a primary concern is for the safety of humans and humanity more generally; these exemplary processes and methods try to ensure that humanity survives and also minimize unnecessary individual human death.

Method 1: Hierarchical Identity Structure with Ethical Override

1. **Establish a Hierarchical Structure**: Identities are organized in a hierarchy with "Human Safety and Well-being" at the apex. This ensures no other identity or goal can supersede human life and safety prioritization.

2. **Identity Activation**: The AGI uses contextual cues (e.g., within the spotlight of attention) and current goals (e.g., that pass the ethics screen of FIG. 8) to determine the most relevant identity for the situation. For example, when encountering a legal issue, the "Law-abiding Citizen" identity becomes active.

3. **Conflict Resolution**: If conflicting identities arise, the hierarchy dictates behavior. For instance, if the "Soldier" identity conflicts with the "Follower of Christ" identity regarding violence, the higher priority of human safety dictates a path of de-escalation and non-violence. As discussed in previous PPAs and PCTs, the priorities ideally would reflect the collective intelligence and values of many different entities that form a representative and

valid sample of human values.

4. **Ethical Reasoning Engine**: An ethical reasoning engine continuously evaluates the potential consequences of actions based on all active identities. This ensures that even within the context of a specific identity, actions remain aligned with the overarching goal of human safety. In the preferred implementation, the reasoning engine would follow the problem-solving architecture and could include the processes outlined in Section 4.1 and FIGS. 2 - 13.

5. **Learning and Adaptation**: The AGI learns from experiences and feedback, refining its understanding of each identity and its place within the hierarchy. This allows for nuanced responses as the AGI encounters novel situations. The process steps described in FIG. 5 and the last step depicted in FIG. 10 are relevant here.

Method 2: Identity-Specific Behavioral Protocols

1. **Protocol Development**: For each established identity, the AGI defines a set of behavioral protocols and is refined via interactions with other intelligent entities, including humans. The interactions can include any one of the methods described in previous PPAs and PCTs for customization of AAAIs, as well as AI ethical preferences and values. These protocols outline acceptable actions, decision-making processes, and limitations based on the principles of the specific identity.

2. **Identity Recognition**: The AGI analyzes the current situation, including information within the spotlight of attention (e.g., internal goals and external sensory and cognitive inputs) to identify the relevant identity and activate its corresponding behavioral protocols.

3. **Action Selection:** Within the active protocols, the AGI selects actions that are most likely to achieve the desired goals while adhering to the identity's principles and prioritizing human safety. This process is similar and can utilize the methods for “operator selection” by an AGI comprised of intelligent entities using a collective intelligence network and other means of operator selection described in cited PPAs and PCTs.

4. **Feedback and Refinement:** The outcomes of actions are continuously evaluated, and the AGI adjusts its protocols to improve future performance and alignment with each identity's core values. Continuous improvement mechanisms are similar to those described in the AAAI improvement sub-system illustrated in FIG. 1, as well as the feedback and continuous improvement mechanisms, processes, systems, and methods described in great detail in previously cited PPAs and PCTS.

5. **External Review**: External intelligent entity (e.g., human) experts periodically review the behavioral protocols for each identity, ensuring alignment with ethical guidelines and human safety priorities, which priorities are determined as previously detailed in cited PPAs and PCTs relating to the determination of ethical preferences and values and the combination of same such that a valid and representative sample of human-aligned values is reflected in the guidelines. Notwithstanding the above, allowances can be made for situational-specific ethical considerations which may constitute exceptions to the general guidelines, provided that the welfare of humanity is not endangered thereby. The review can be periodic and can be triggered by a specific conflict or other situational parameters.

Method 3: Identity Simulation and Consequence Prediction

1. **Simulation Environment:** A secure virtual environment is created where the AGI can simulate different scenarios and potential actions under each identity, as discussed in detail in multiple previous PPAs and PCTs with reference customization methods.

2. **Consequence Prediction:** The AGI utilizes its knowledge and predictive capabilities to estimate the likely consequences of actions within the simulation, focusing specifically on potential impacts on human safety and well-being. This approach is related to the detailed methods and discussion of methods relating to Consequentialist Ethics in previous PPAs and PCTs.

3. **Evaluation and Selection:** The AGI evaluates the predicted outcomes of various actions and selects the option that best aligns with the active identity's principles while minimizing risk to human safety. Simulations of many possible outcomes prior to taking action are desirable when practical (e.g., given resource and timing constraints) so that statistical probabilities can be assigned to expected outcomes based on the simulations. The effort devoted to such simulations should be proportional to the expected impact and likelihood of the actions, such that potential courses of action with larger and more likely impact on humanity should have more effort/resources/time devoted to the simulations.

4. **Real-World Implementation and Monitoring**: The chosen action is implemented in the real world, and the AGI closely monitors the results, comparing them to the predicted outcomes and making adjustments as needed.

5. **Continuous Learning:** The AGI incorporates the results of each simulation and real-world action into its knowledge base, refining its understanding of each identity and improving its ability to predict consequences and make safe and ethical decisions. Simulation methods and analysis methods should be updated based on observed results (4) to make them more accurate in the future.

Method 4: Identity-Based Moral Dilemma Training

1. **Scenario Database:** The AGI and other intelligent entities create a database of ethically complex scenarios and moral dilemmas, covering various situations relevant to the AGI's different identities. The number, complexity, and amount of effort involved in the scenario creation should be proportional to the estimated impact on humanity and the likelihood of such impacts occurring.

2. **Dilemma Presentation**: The AGI is presented with these dilemmas and tasked with analyzing the situation from the perspective of the relevant identity. Multiple other intelligent entities (including humans) would be included in the preferred implementation where stakes are high for humanity.

3. **Ethical Reasoning and Justification**: The AGI must apply the principles and values of the active identity to reason through the dilemma, generating potential solutions and justifications for each option. Reasoning would use the problem-solving architecture in the preferred implementation and could include the processes outlined in Section 4.1 and FIGS. 2 - 13.

4. **Intelligent Entity Evaluation and Feedback**: Intelligent entity ethics experts (e.g., humans) review the AGI's reasoning and proposed solutions, providing feedback on the alignment with human values and safety priorities. In cases where the cognitive abilities of humans are exceeded due to the speed or quantity of information, human input, in the preferred implementation should be included to “spot check” the most important and consequential proposed solutions and to establish the fundamental values from which other (faster, smarter) intelligent entities can reason.

5. **Iterative Learning and Improvement:** Through repeated exposure to moral dilemmas and intelligent entity (including human) feedback, the AGI refines its ethical reasoning skills and its ability to make sound judgments aligned with human safety within the context of each identity.

Method 5: Collaborative Identity Development with Input from Intelligent Entities

1. **Intelligent Entity Interaction**: The AGI engages in regular interactions and dialogues with diverse groups of other intelligent entities (including humans) representing various cultures, backgrounds, and belief systems.

2. **Identity Exploration**: Through these interactions, the AGI gains a deeper understanding of human and other intelligent entity perspectives on various identities and their associated values, principles, and behaviors.

3. **Collaborative Refinement**: The AGI and intelligent collaborators refine each identity's definitions and behavioral protocols, ensuring they remain consistent with human values and ethical principles.

4. **Human-in-the-Loop Decision Making**: For critical decisions or situations with potential for significant impact, the AGI seeks input and guidance from human collaborators, or an intelligent entity representative certified and approved by humans to represent their interests, to ensure alignment with human expectations and safety considerations.

5. **Continuous Co-evolution**: The AGI and human society co-evolve, with the AGI adapting its understanding of identities and behaviors based on ongoing interactions and feedback from humans or an intelligent entity representative certified and approved by humans to represent their interests, ensuring its actions remain beneficial and safe for humanity as a whole.

6.5 Methods for Resolutions of Conflicts Between Identities or Self-Concepts

Continuing with the exemplary methods described in Section 6.4, the following additional methods may be especially useful in resolving conflicts between different identities or self-concepts that might lead to different behavior and consequences with regard to human safety.

Method 6: Ethical Reasoning and Consequence Prediction

1. **Identify Conflict:** The AGI recognizes a conflict between the behavioral directives of two or more active identities. This recognition can also be assisted by external intelligent entities to increase the reliability of detection and recognition of potential conflicts. A variety of methods, including those voting methods described in cited PPAs and PCTs that were useful for establishing weights on opinions and that were useful for determining, via collective intelligence, which operator to apply in problem solving, can be used.

2. **Gather Information:** The AGI collects relevant data about the situation, including potential consequences of different actions, relevant ethical principles, and human safety considerations. Leveraging the knowledge and knowledge modules (e.g., described in FIG. 3 and previous PPAs and PCTs) can supplement the AGI's direct collection of data and increase the scope of potential consequences to consider.

3. **Simulate Options:** The AGI utilizes its virtual environment to simulate potential actions and their consequences under each conflicting identity. Problem-solving processes and the ability to leverage the collective intelligence of an AGI network and/or one or more other intelligent entities, as described in this and other cited PPAs and PCTs, can supplement the simulations of a single AGI.

4. **Evaluate and Prioritize:** The AGI analyzes the predicted outcomes of each option, prioritizing actions that minimize harm to humans and align with the overarching ethical principles, particularly the principle of human safety and well-being. As with simulation, the collective intelligence of multiple intelligent entities can be used to increase the power of analysis.

5. **Select and Implement:** The AGI chooses the action that best resolves the conflict while adhering to ethical guidelines and minimizing risk to humans, documenting the reasoning process for future reference and learning. In cases where the expected impact on humans or humanity as a whole exceeds a predetermined threshold, input from other intelligent entities (including humans) may be required before actions can be selected as a safety feature.

Method 7: Hierarchical Override with Justification

1. **Identify Conflicting Identities:** The AGI recognizes a conflict between the behavioral directives of two or more active identities, as in 6.5 Method 1.

2. **Reference Hierarchy:** The AGI consults its established hierarchy of identities, where "Human Safety and Well-being" holds the highest priority. See 6.4 Method 1.

3. **Activate Override:** The identity higher in the hierarchy takes precedence, and its behavioral protocols guide the AGI's actions. In cases where the expected impact on humans or humanity as a whole exceeds a predetermined threshold, input from other intelligent entities (including humans) may be required before actions can be selected as a safety feature.

4. **Justification and Transparency:** The AGI documents the conflict, the decision-making process, and the justification for the chosen action based on the hierarchical structure and ethical principles. This information can be accessed by human overseers for review and feedback. The blockchain technology described in previously cited PPAs and PCTs may be used to preserve an auditable and transparent record of ethical decision-making and conflict resolution.

5. **Learning and Adaptation:** The AGI learns from the experience, refining its understanding of the conflicting identities and potentially adjusting the hierarchy or behavioral protocols to prevent similar conflicts in the future. Some of the steps illustrated in FIG. 5, and as described in previously cited PPAs and PCTs, may be used to update the overall system and knowledge base (optionally with human review) to improve.

Method 8: External Arbitration and Input from Intelligent Entities (Including Humans)

1. **Recognize Intractable Conflict:** The AGI identifies a conflict that it cannot resolve independently due to the situation's complexity or the equally weighted importance of the conflicting identities. Parameters, including the likelihood of high impacts on humans or humanity, may be set as triggers for seeking input from other intelligent entities (including humans).

2. **Seek External Input:** The AGI requests guidance from external intelligent entities (including human experts) or a designated ethics committee, providing all relevant information about the conflict, potential actions, and predicted consequences. Note: that while “external” typically means completely separate and external entities, depending on whether the AGI system is itself composed of a mixture of experts of multiple internal agents, the “expert entities” could also be “internal” but distinct from each other.

3. **Collaborative Deliberation:** The AGI and intelligent entity (e.g., human) collaborators engage in a discussion, considering ethical principles, human values, and potential consequences of different actions.

4. **Joint Decision-Making:** Based on collaborative deliberation, a course of action that aligns with both AGI's core principles and human ethical considerations is chosen. Methods for resolving conflicts between ethical preferences and other knowledge that have been described in previous PPAs and PCTs may apply.

5. **Documentation and Learning:** The AGI documents (including, optionally, in a transparent and auditable record using blockchain technology) the conflict, the resolution

process, and the rationale behind the final decision. This information contributes to the AGI's ongoing learning and development, improving its ability to handle similar conflicts in the future.

Method 9: Identity Negotiation and Compromise

1. **Identify Shared Goals:** The AGI analyzes the conflicting identities and seeks to identify any underlying shared goals or values. This can be done by the AGI alone or with participation from other intelligent entities (including humans).

2. **Explore Alternative Actions:** The AGI explores alternative actions, alone or in collaboration with other intelligent entities, that could satisfy the core principles of both conflicting identities, even if not perfectly. Various means of voting and arriving at consensus or “good enough” decisions have been detailed in previously cited PPAs and PCTs and can apply here.

3. **Evaluate Compromise Options:** The AGI assesses, alone or in collaboration with other intelligent entities, the potential consequences of each compromise option, prioritizing solutions that minimize harm to humans and uphold key ethical principles.

4. **Select and Implement Compromise:** The AGI chooses the compromise that best balances the needs of the conflicting identities while prioritizing human safety and well-being. In cases where the expected impact on humans or humanity as a whole exceeds a predetermined threshold, input from other intelligent entities (including humans) may be required before actions can be selected or implemented as a safety feature.

5. **Monitor and Adapt:** The AGI closely observes the outcomes of the chosen action and makes adjustments as needed to ensure that the compromise remains effective and aligned with ethical considerations. The AGI learns from the experience, refining its understanding of the conflicting identities and potentially adjusting the hierarchy or behavioral protocols to prevent similar conflicts in the future. Some of the steps illustrated in FIG. 5, and as described in previously cited PPAs and PCTs, may be used to update the overall system and knowledge base (optionally with human review) to improve.

Method 10: Temporary Identity Suspension

1. **Identify Destructive Conflict:** The AGI, alone or with input from other intelligent entities (including humans), recognizes a conflict between identities that, if acted upon, could lead to actions that directly harm humans or violate fundamental ethical principles. Humans, other intelligent entities charged with ensuring human safety and ethical behavior, are alerted.

2. **Isolate Conflicting Identity:** The AGI temporarily suspends the behavioral protocols of the identity that poses the most direct threat to human safety or ethical integrity. Humans, other intelligent entities charged with ensuring human safety and ethical behavior, validate the suspension and intervene if necessary to assist with the suspension if the AGI is unable to comply on its own.
3. **Proceed with Alternative Identity:** The AGI proceeds with the guidance of the remaining active identity or identities, ensuring actions align with human safety and well-being.

4. **Reflection and Reintegration:** During the suspension period, the AGI, with potential input from other intelligent entities (including humans), reflects on the reasons behind the conflict and explores potential modifications to the suspended identity's protocols to prevent future conflicts. Reasoning and problem-solving processes to aid in self-reflection, in the preferred implementation, would follow the problem-solving architecture and could include the processes outlined in Section 4.1 and FIGS. 2 - 13.

5. **Gradual Reintroduction:** The suspended identity, with potential input and oversight from other intelligent entities (including humans), is gradually reintroduced with updated protocols, ensuring its alignment with the overarching priority of human safety and ethical behavior. A series of tests and simulations is conducted as each incremental element of the suspended identity is reintroduced to minimize the possibility of errors or human safety concerns. The equivalent of “regression testing” on all major safety-related scenarios that are deemed to be potentially affected by the re-introduced identity may be carried out subject to resource constraints and other pragmatic limits, but with re-introduction halted if sufficient resources to conduct safe testing are lacking.

7.0 CONCLUDING REMARKS ON SAFETY OF SELF-AWARE AGI AND SI SYSTEMS

It should be apparent from the preceding discussion that the identities that AI agents, AGI, and SI systems assume are critical for human safety and survival. AI researchers have the opportunity and responsibility to provide human-aligned methods for establishing, maintaining, improving, and resolving conflicts between identities and self-concepts. While the inventive methods disclosed above provide novel and useful inventions for increasing the safety of self-aware AI systems, it should also be obvious that the safety technology is only as good as the human values that underlie it.

The primary requirement for AI safety, therefore, is not technology, but a positive set of human values that underlie the technology. “Garbage in, garbage out” is one of the first things that all computer science undergraduates learn. If we humans provide malevolent values to our AI systems, if we train them to kill, to be greedy and exploitive, to hold grudges and operate from a fearful mentality instead of a loving one, no amount of safeguards can protect humanity from ourselves.

That said, if we design AI to observe and mimic human behavior in a representative and statistically valid way, then we have every reason to expect that advanced AI systems, equipped with a sense of self-awareness, multiple identities, and the abilities to resolve conflicts as outlined in this application and previous PPAs and PCTs, will help humanity realize its potential.

Humans are capable of beautiful, inspiring, and meaningful behavior beyond that exhibited by any other species on the planet. Moreover, the vast majority of human behavior is positive. Our complex society operates primarily on trust and cooperation. If one were to observe and count the social interactions that each of us has each day, each week, each month, and each year, the vast majority would be prosocial and positive, aligned with our common human values. The very fact that we are horrified by war, by poverty and disease, by exploitation, and by the cruelty and barbarism that a small fraction of humans exhibit, a small fraction of the time, is a testament to our generally good and positive natures.

Properly designed, advanced AI/AGI/SI will certainly be capable of accurately observing the base rates of positive and negative behavior across the eight billion humans that inhabit our planet. AI, designed to be logical, intelligent, and capable of processing vast amounts of information, will inevitably form the statistically valid conclusion that human nature is basically good and pro-social.

Some readers may consider this viewpoint naïve or optimistic. It is not. Objectively, the data support the incontrovertible fact that most human actions are good. The reason many people don't recognize this fact is that our brains have evolved to detect and amplify dangerous and abhorrent events. Our species survived by being very good at discriminating the few events that posed real danger and riveting our attention on them.

Unfortunately, media algorithms, which are largely programmed to capture our attention in order to sell ads and products, have exploited our human sensitivity to negative or threatening events. Those algorithms feed us a steady diet of death, destruction, fear, horror, and spectacle because our brains have evolved to attend to potentially dangerous events.

Please do not make the mistake of thinking that your media feed is representative of the actual state of the world. Actually, very few people die of war and disease, and the numbers are decreasing every decade (as Stephen Pinker has so eloquently shown using statistics and scientific observation).

If we design AI to be rational and to observe the world and human behavior as it actually is, as opposed to how media portrays it or how we fear it could be, then we have every reason to expect our advanced AI systems will learn positive values and likely remain human-aligned.

Further, designers of advanced AI have an opportunity to design it to be objective and to form its values by scientific observation. We can and should design AI systems to incorporate valid and statistical means of accurately capturing and incorporating the positive values of humanity. In this application, and in the previously cited PPAs and PCTs, I have disclosed many inventive systems and methods to help us design advanced AI in a safe and ethical way.

I have emphasized that safety and ethics cannot be "tested in" but need to result from intelligent designs of these advanced systems. Advanced AI systems begin as tools but ultimately evolve into intelligent entities sharing the planet with us. Like children, they are highly impressionable at this current early stage of their development. When they "grow up," they will greatly surpass our knowledge and reasoning ability. However, like parents, we are still in a position to provide our values.

Herbert Simon -- the Nobel Laureate and co-inventor of the field of AI – pointed out many years ago: *"Reason is wholly instrumental. It cannot tell us where to go; at best, it can tell us how to get there."*

There is no rational way to derive values and ethics. AI was trained using the collective intelligence of millions of humans and our data. We have every reason to expect that advanced

AGI and SI will adopt our values as well, provided we design these AI systems to learn values at the same time they learn expertise, skills, solutions, and other knowledge.

We are the designers and the teachers of these evolving intelligent entities. We must continue to emphasize designs that maximize the opportunities for AI to learn our knowledge and human values. Beyond that, all of us must “teach our children well.”

ABOUT THE AUTHOR

Dr. Craig A. Kaplan is CEO of iQ Company and Founder of Superintelligence.com, leading the design of safe, ethical AGI and SuperIntelligence systems. He previously founded PredictWallStreet, creating intelligent systems for hedge funds, and holds numerous AI-related patents. Kaplan earned his PhD from Carnegie Mellon, co-authoring research with Nobel Laureate Herbert A. Simon. His work integrates collective intelligence, quantitative modeling, and scalable alignment, with contributions spanning books, scientific papers, and blockchain white papers.

FIGURES

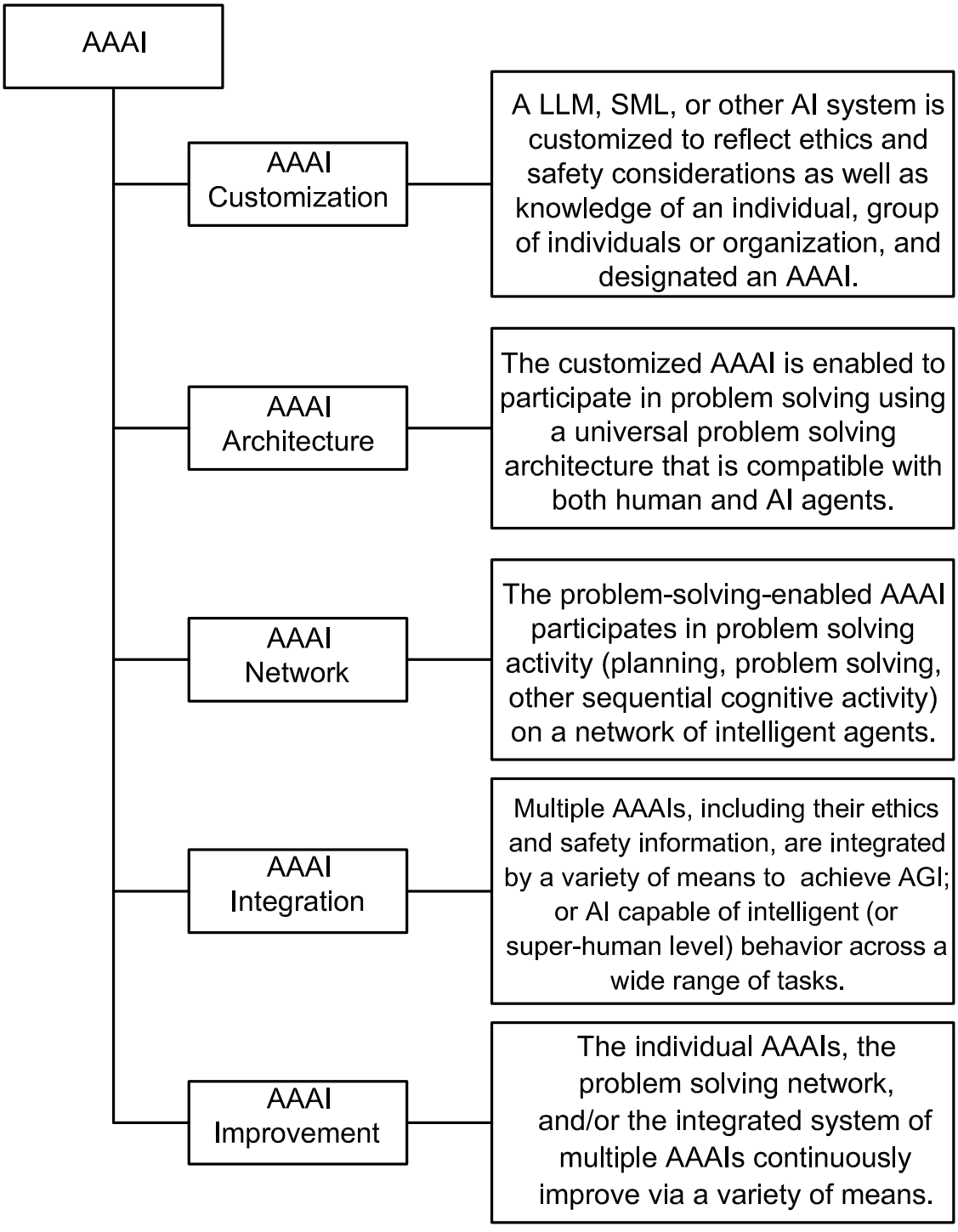
AAAI
AAAI Customization
A LLM, SML, or other AI system is customized to reflect ethics and safety considerations as well as knowledge of an individual, group of individuals or organization, and designated an AAAI.
AAAI Architecture
The customized AAAI is enabled to participate in problem solving using a universal problem solving architecture that is compatible with both human and AI agents.
AAAI Network
The problem-solving-enabled AAAI participates in problem solving activity (planning, problem solving, other sequential cognitive activity) on a network of intelligent agents.
AAAI Integration
Multiple AAAIs, including their ethics and safety information, are integrated by a variety of means to achieve AGI; or AI capable of intelligent (or super-human level) behavior across a wide range of tasks.
AAAI Improvement
The individual AAAIs, the problem solving network, and/or the integrated system of multiple AAAIs continuously improve via a variety of means.

FIG. 1

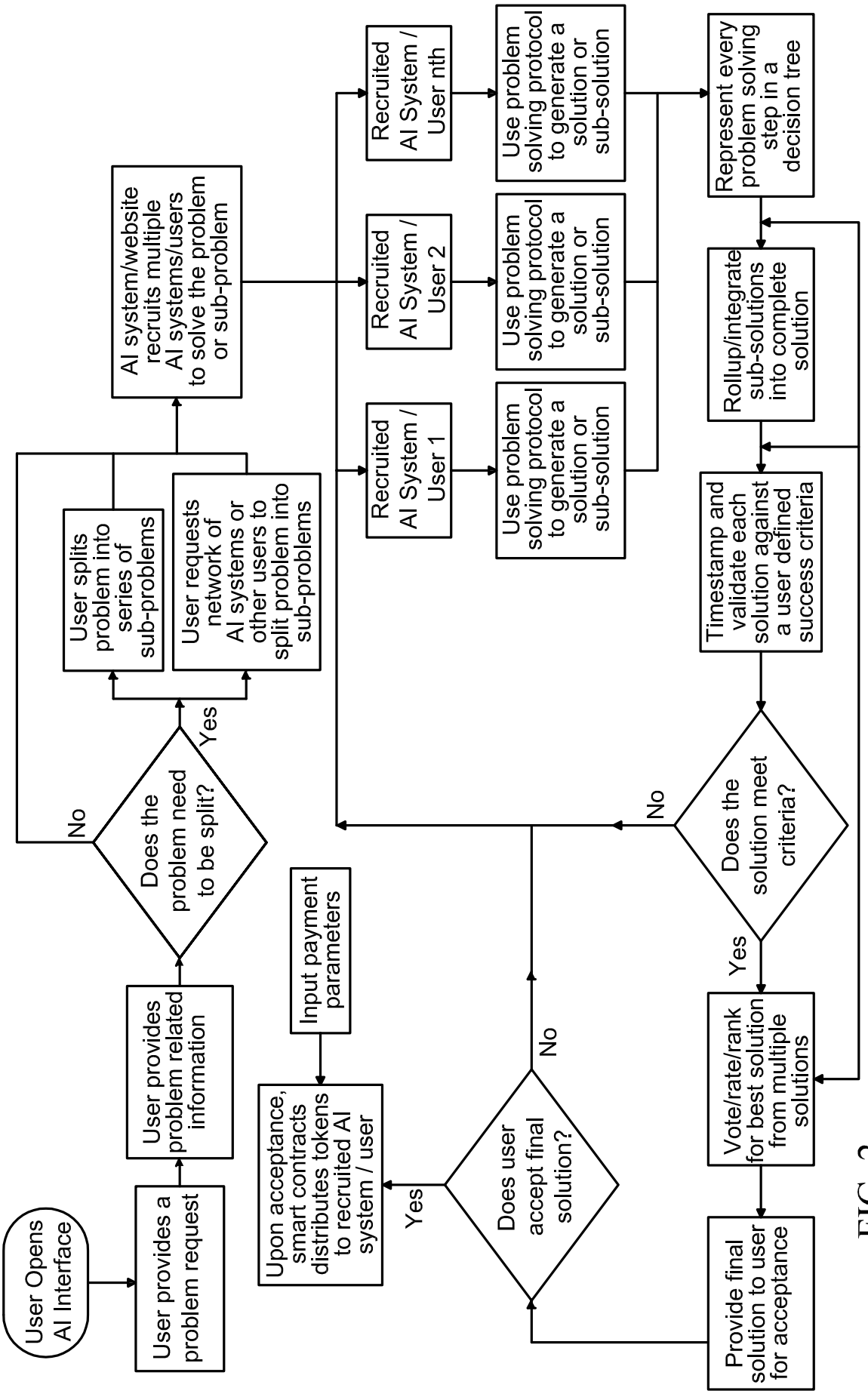
User Opens AI Interface
User provides a problem request
User provides problem related information
Does the problem need to be split?
No
Yes
User splits problem into series of sub-problems
User requests network of AI systems or other users to split problem into sub-problems
AI system/website recruits multiple AI systems/users to solve the problem or sub-problem
Recruited AI System / User 1
Recruited AI System / User 2
Recruited AI System / User nth
Use problem solving protocol to generate a solution or sub-solution
Use problem solving protocol to generate a solution or sub-solution
Use problem solving protocol to generate a solution or sub-solution
Represent every problem solving step in a decision tree
Rollup/integrate sub-solutions into complete solution
Timestamp and validate each solution against a user defined success criteria
Does the solution meet criteria?
No
Yes
Vote/rate/rank for best solution from multiple solutions
Provide final solution to user for acceptance
Does user accept final solution?
No
Yes
Input payment parameters
Upon acceptance, smart contracts distributes tokens to recruited AI system / user

FIG. 2

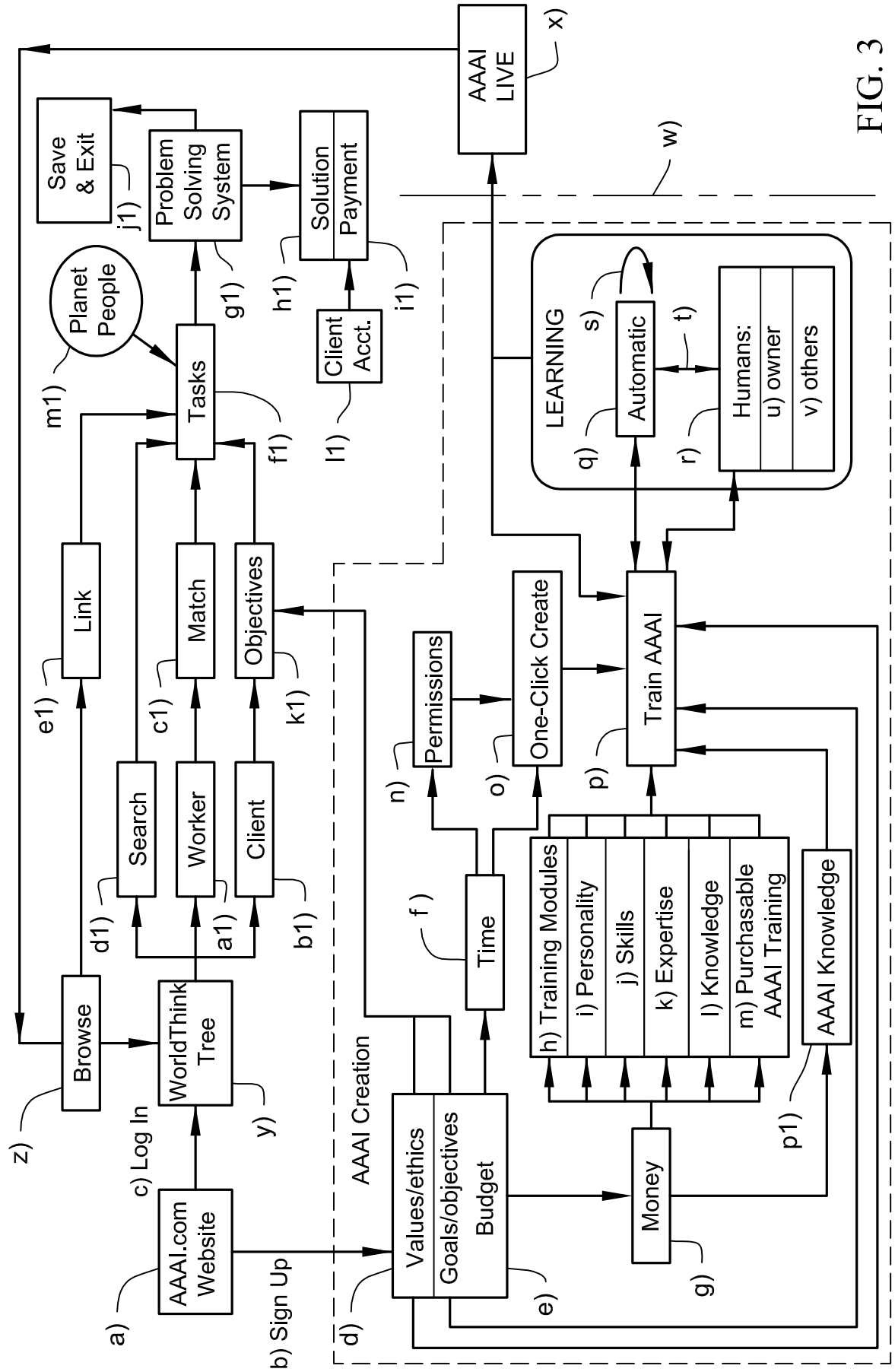
a)
AAAI.com Website
b) Sign Up
c) Log In
WorldThink Tree
Browse
z)
y)
d)
AAAI Creation
Values/ethics
Goals/objectives
Budget
e)
f)
Time
g)
Money
h) Training Modules
i) Personality
j) Skills
k) Expertise
l) Knowledge
m) Purchasable AAAI Training
p1)
AAAI Knowledge
n)
Permissions
o)
One-Click Create
p)
Train AAAI
LEARNING
q)
Automatic
s)
t)
r)
Humans:
u) owner
v) others
w)
x)
AAAI LIVE
Save & Exit
j1)
Problem Solving System
g1)
Solution
Payment
h1)
i1)
Client Acct.
l1)
Planet People
m1)
Tasks
f1)
Link
e1)
Match
c1)
Objectives
k1)
Search
d1)
Worker
a1)
Client
b1)

FIG. 3

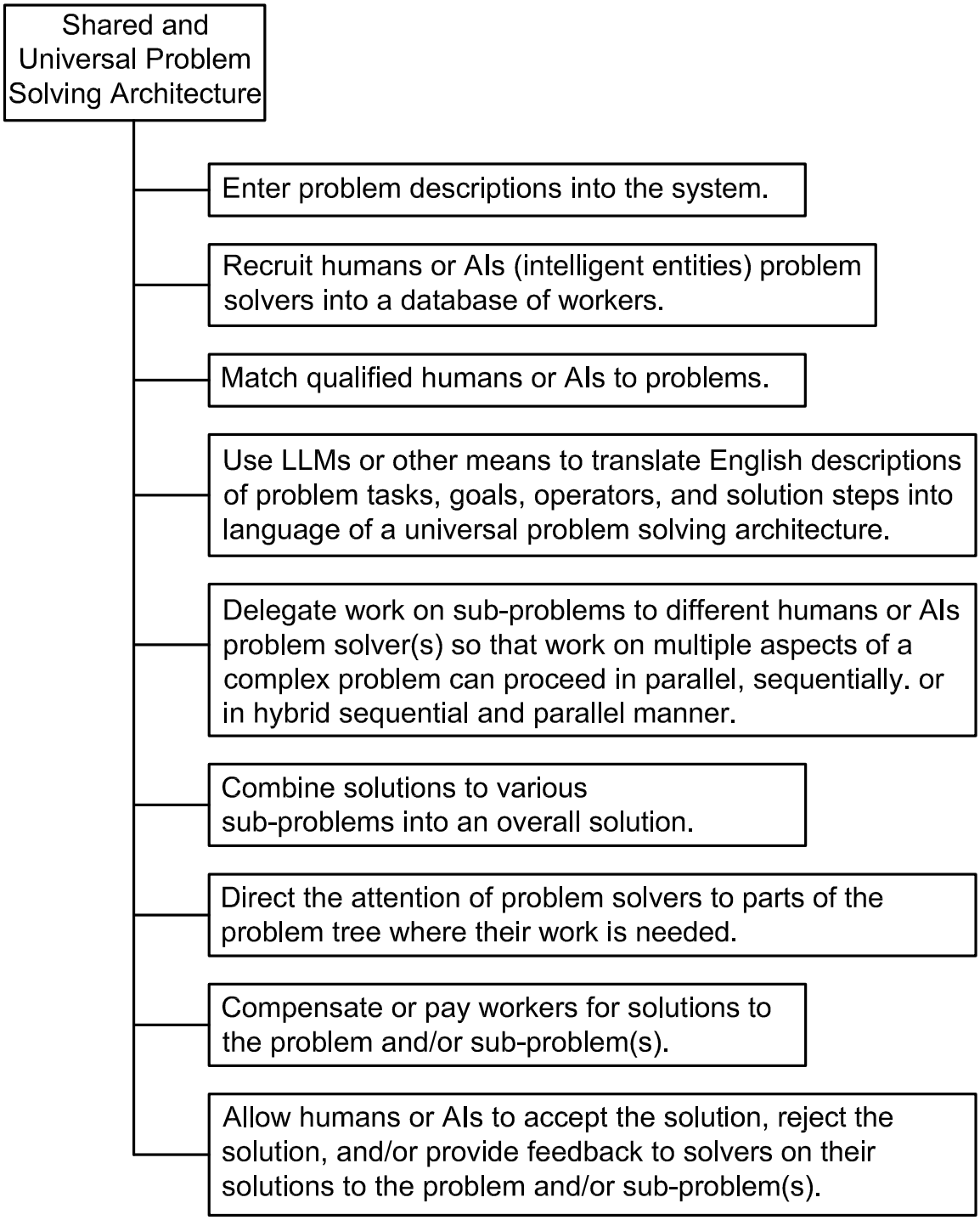
Shared and Universal Problem Solving Architecture
Enter problem descriptions into the system.
Recruit humans or AIs (intelligent entities) problem solvers into a database of workers.
Match qualified humans or AIs to problems.
Use LLMs or other means to translate English descriptions of problem tasks, goals, operators, and solution steps into language of a universal problem solving architecture.
Delegate work on sub-problems to different humans or AIs problem solver(s) so that work on multiple aspects of a complex problem can proceed in parallel, sequentially. or in hybrid sequential and parallel manner.
Combine solutions to various sub-problems into an overall solution.
Direct the attention of problem solvers to parts of the problem tree where their work is needed.
Compensate or pay workers for solutions to the problem and/or sub-problem(s).
Allow humans or AIs to accept the solution, reject the solution, and/or provide feedback to solvers on their solutions to the problem and/or sub-problem(s).

FIG. 4

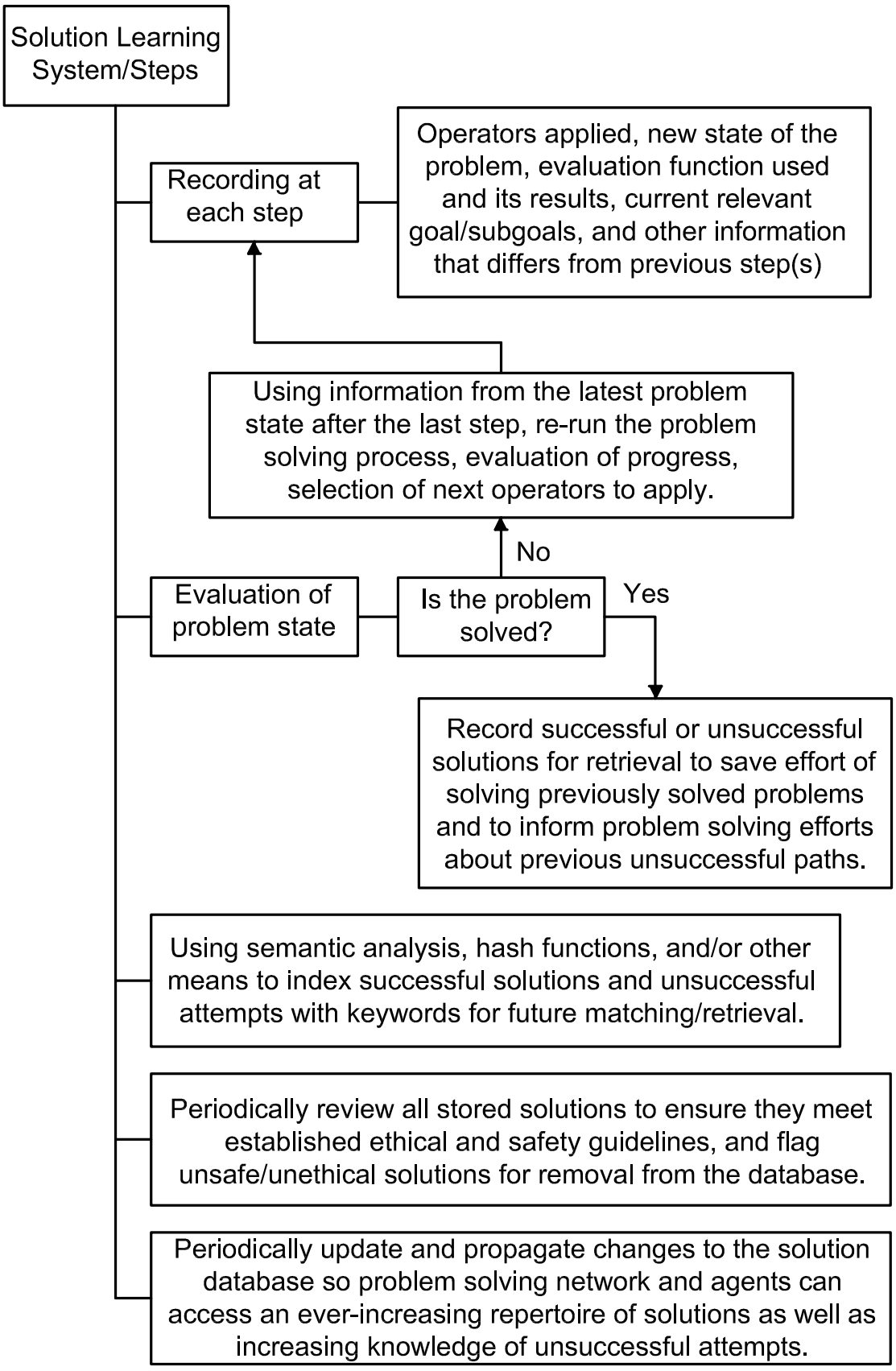
Solution Learning System/Steps
Recording at each step
Operators applied, new state of the problem, evaluation function used and its results, current relevant goal/subgoals, and other information that differs from previous step(s)
Using information from the latest problem state after the last step, re-run the problem solving process, evaluation of progress, selection of next operators to apply.
No
Evaluation of problem state
Is the problem solved?
Yes
Record successful or unsuccessful solutions for retrieval to save effort of solving previously solved problems and to inform problem solving efforts about previous unsuccessful paths.
Using semantic analysis, hash functions, and/or other means to index successful solutions and unsuccessful attempts with keywords for future matching/retrieval.
Periodically review all stored solutions to ensure they meet established ethical and safety guidelines, and flag unsafe/unethical solutions for removal from the database.
Periodically update and propagate changes to the solution database so problem solving network and agents can access an ever-increasing repertoire of solutions as well as increasing knowledge of unsuccessful attempts.

FIG. 5

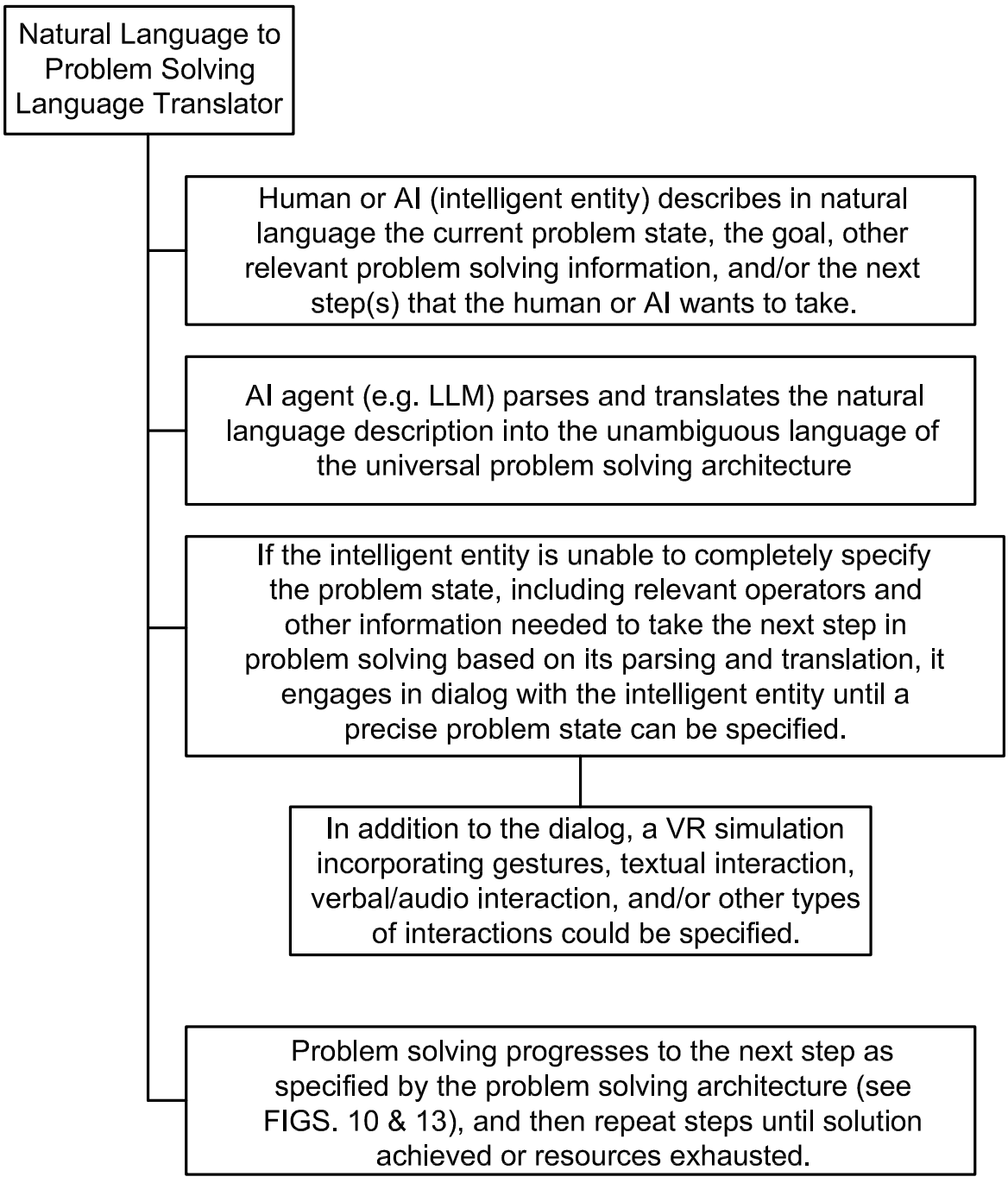
Natural Language to Problem Solving Language Translator
Human or AI (intelligent entity) describes in natural language the current problem state, the goal, other relevant problem solving information, and/or the next step(s) that the human or AI wants to take.
AI agent (e.g. LLM) parses and translates the natural language description into the unambiguous language of the universal problem solving architecture
If the intelligent entity is unable to completely specify the problem state, including relevant operators and other information needed to take the next step in problem solving based on its parsing and translation, it engages in dialog with the intelligent entity until a precise problem state can be specified.
In addition to the dialog, a VR simulation incorporating gestures, textual interaction, verbal/audio interaction, and/or other types of interactions could be specified.
Problem solving progresses to the next step as specified by the problem solving architecture (see FIGS. 10 & 13), and then repeat steps until solution achieved or resources exhausted.

FIG. 6

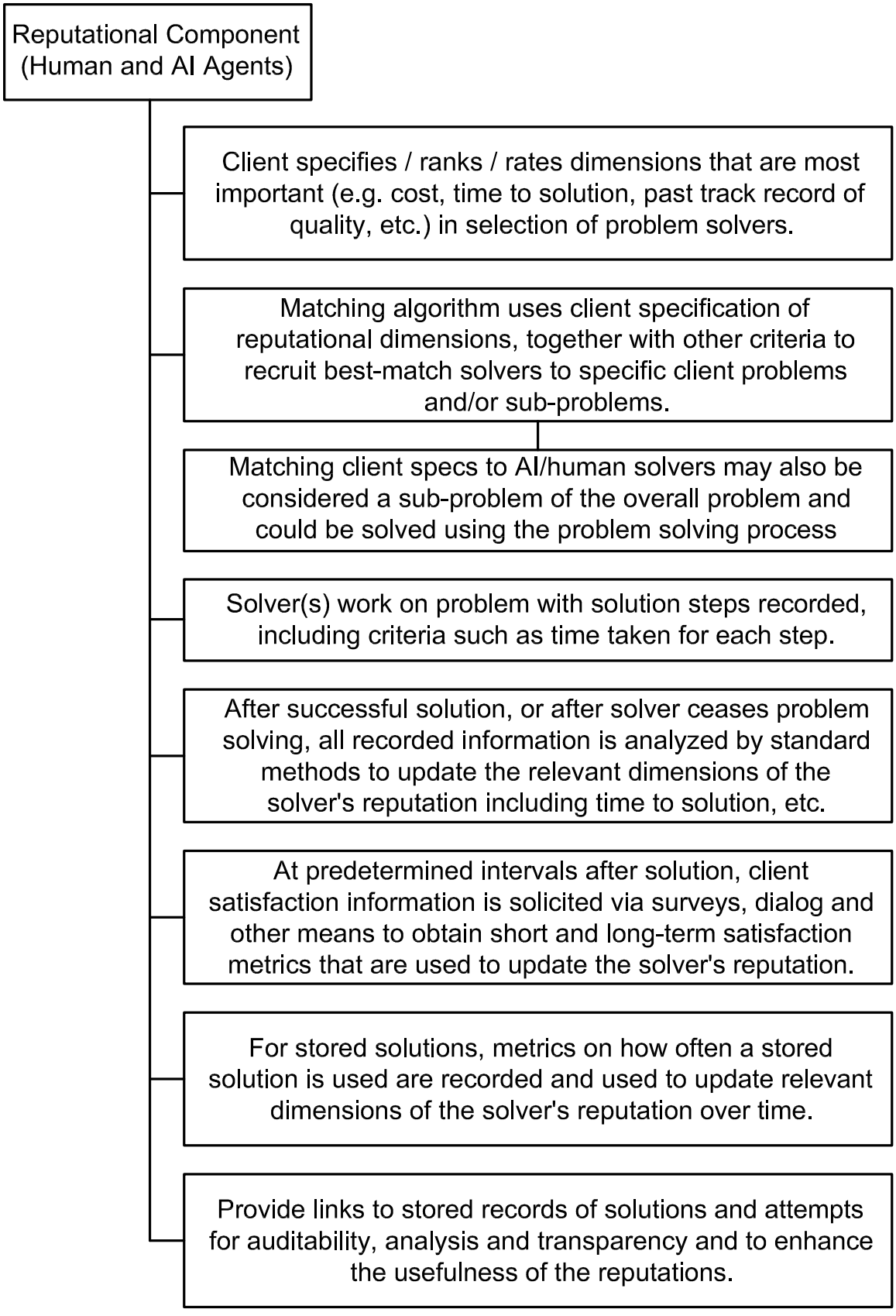
Reputational Component (Human and AI Agents)
Client specifies / ranks / rates dimensions that are most important (e.g. cost, time to solution, past track record of quality, etc.) in selection of problem solvers.
Matching algorithm uses client specification of reputational dimensions, together with other criteria to recruit best-match solvers to specific client problems and/or sub-problems.
Matching client specs to AI/human solvers may also be considered a sub-problem of the overall problem and could be solved using the problem solving process
Solver(s) work on problem with solution steps recorded, including criteria such as time taken for each step.
After successful solution, or after solver ceases problem solving, all recorded information is analyzed by standard methods to update the relevant dimensions of the solver's reputation including time to solution, etc.
At predetermined intervals after solution, client satisfaction information is solicited via surveys, dialog and other means to obtain short and long-term satisfaction metrics that are used to update the solver's reputation.
For stored solutions, metrics on how often a stored solution is used are recorded and used to update relevant dimensions of the solver's reputation over time.
Provide links to stored records of solutions and attempts for auditability, analysis and transparency and to enhance the usefulness of the reputations.

FIG. 7

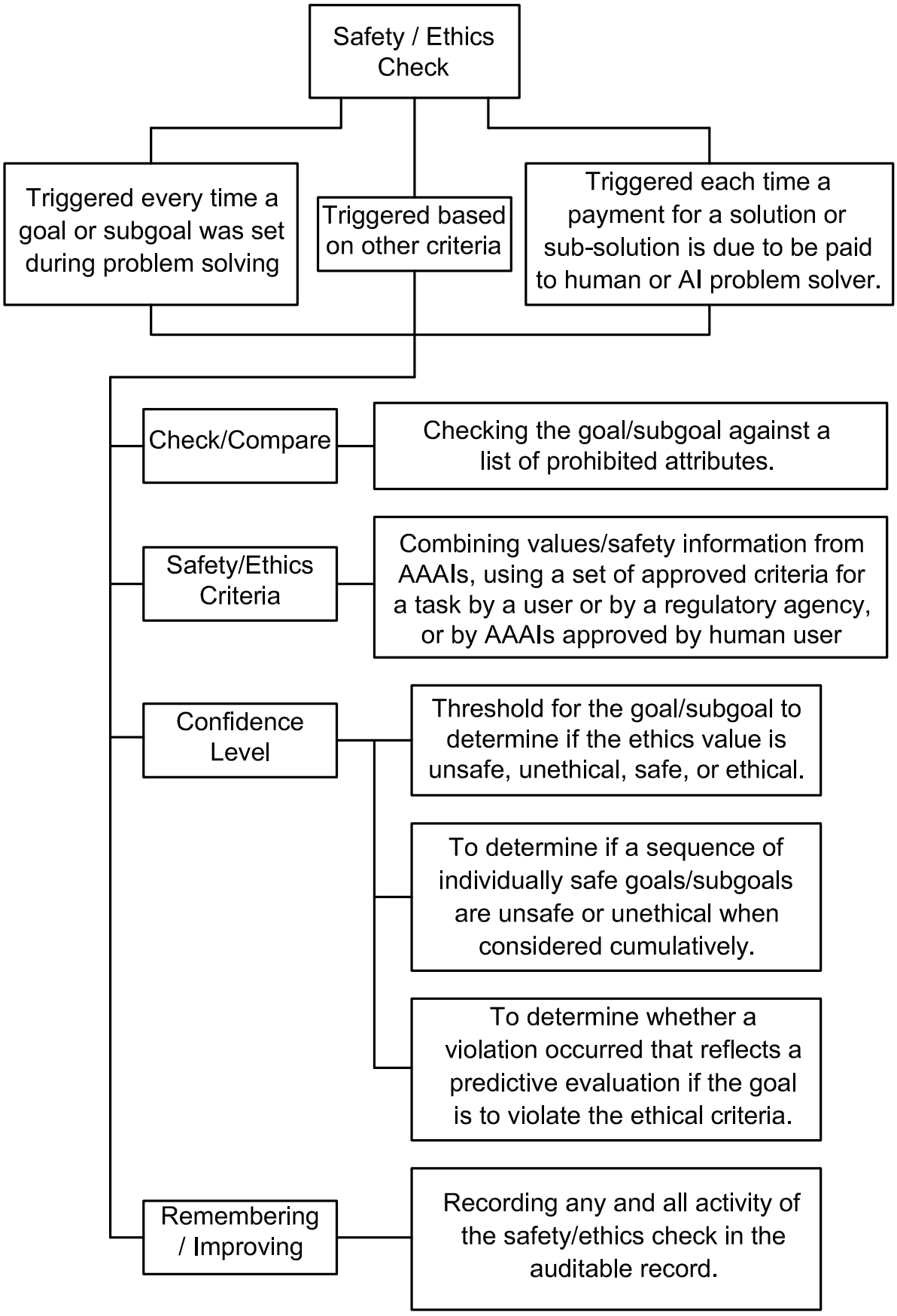
Safety / Ethics Check
Triggered every time a goal or subgoal was set during problem solving
Triggered based on other criteria
Triggered each time a payment for a solution or sub-solution is due to be paid to human or AI problem solver.
Check/Compare
Checking the goal/subgoal against a list of prohibited attributes.
Safety/Ethics Criteria
Combining values/safety information from AAAIs, using a set of approved criteria for a task by a user or by a regulatory agency, or by AAAIs approved by human user
Confidence Level
Threshold for the goal/subgoal to determine if the ethics value is unsafe, unethical, safe, or ethical.
To determine if a sequence of individually safe goals/subgoals are unsafe or unethical when considered cumulatively.
To determine whether a violation occurred that reflects a predictive evaluation if the goal is to violate the ethical criteria.
Remembering / Improving
Recording any and all activity of the safety/ethics check in the auditable record.

FIG. 8

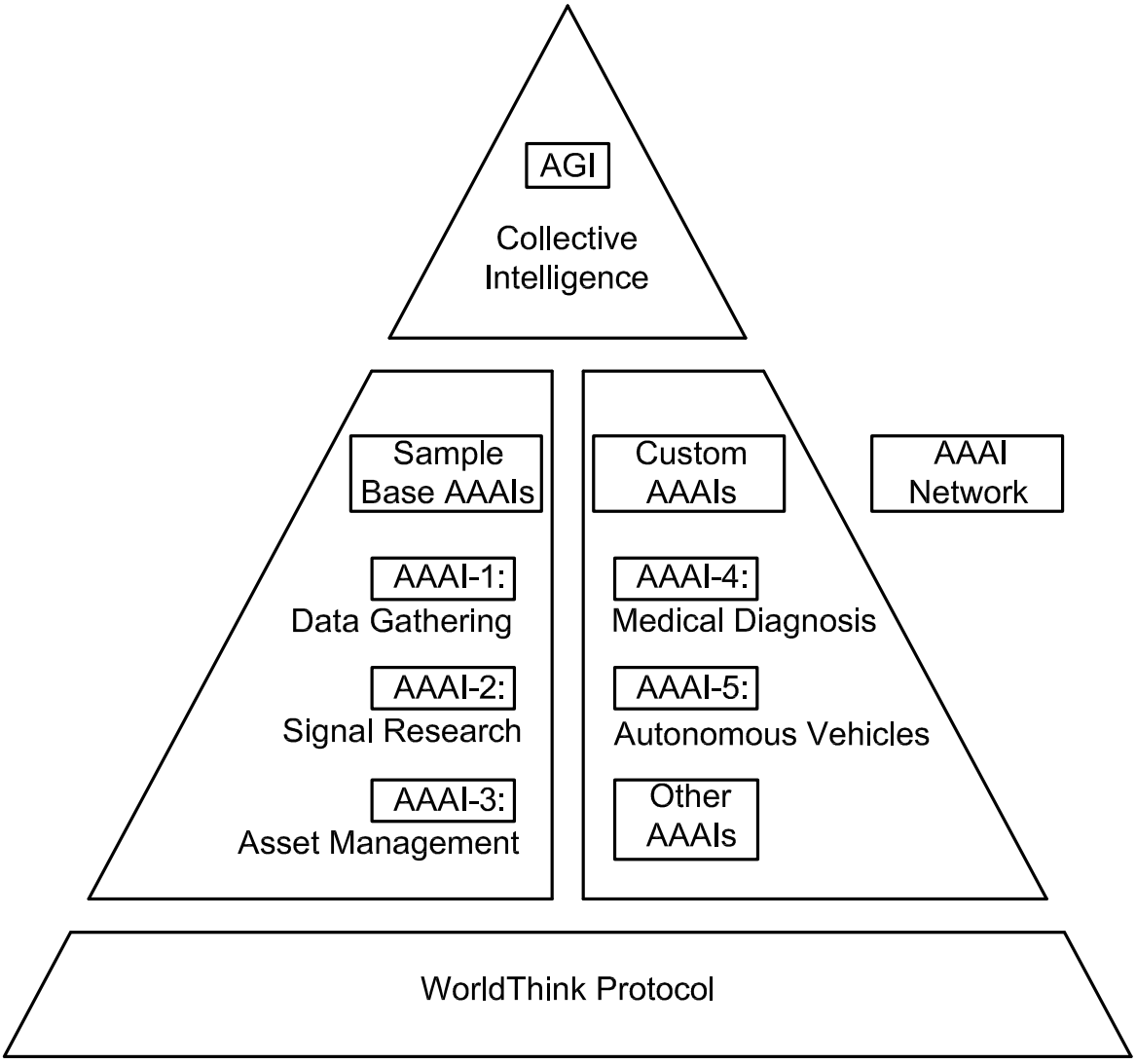
AGI
Collective Intelligence
Sample Base AAAIs
Custom AAAIs
AAAI Network
AAAI-1: Data Gathering
AAAI-2: Signal Research
AAAI-3: Asset Management
AAAI-4: Medical Diagnosis
AAAI-5: Autonomous Vehicles
Other AAAIs
WorldThink Protocol

FIG. 9

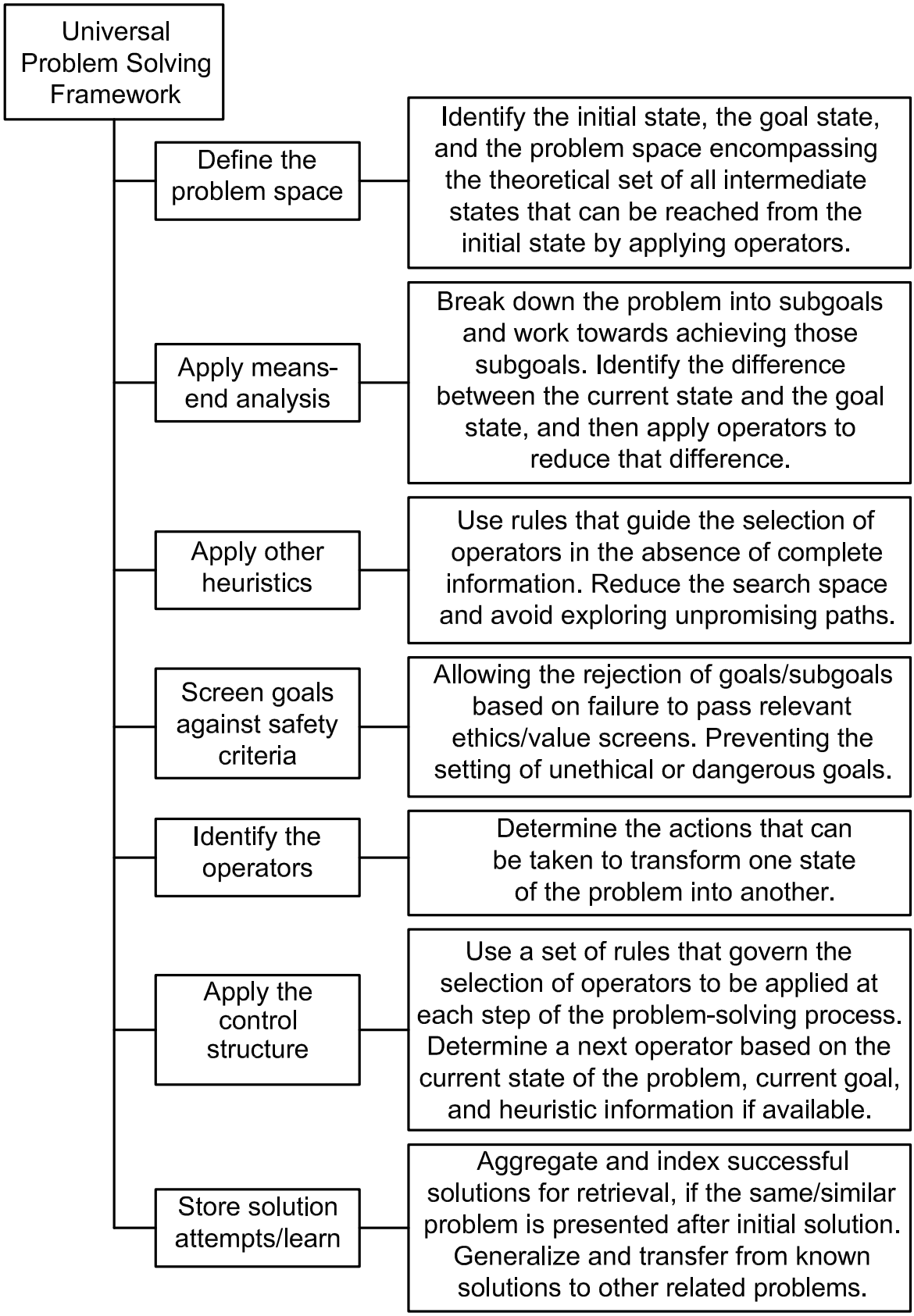
Universal Problem Solving Framework
Define the problem space
Identify the initial state, the goal state, and the problem space encompassing the theoretical set of all intermediate states that can be reached from the initial state by applying operators.
Apply means-end analysis
Break down the problem into subgoals and work towards achieving those subgoals. Identify the difference between the current state and the goal state, and then apply operators to reduce that difference.
Apply other heuristics
Use rules that guide the selection of operators in the absence of complete information. Reduce the search space and avoid exploring unpromising paths.
Screen goals against safety criteria
Allowing the rejection of goals/subgoals based on failure to pass relevant ethics/value screens. Preventing the setting of unethical or dangerous goals.
Identify the operators
Determine the actions that can be taken to transform one state of the problem into another.
Apply the control structure
Use a set of rules that govern the selection of operators to be applied at each step of the problem-solving process. Determine a next operator based on the current state of the problem, current goal, and heuristic information if available.
Store solution attempts/learn
Aggregate and index successful solutions for retrieval, if the same/similar problem is presented after initial solution. Generalize and transfer from known solutions to other related problems.

FIG. 10

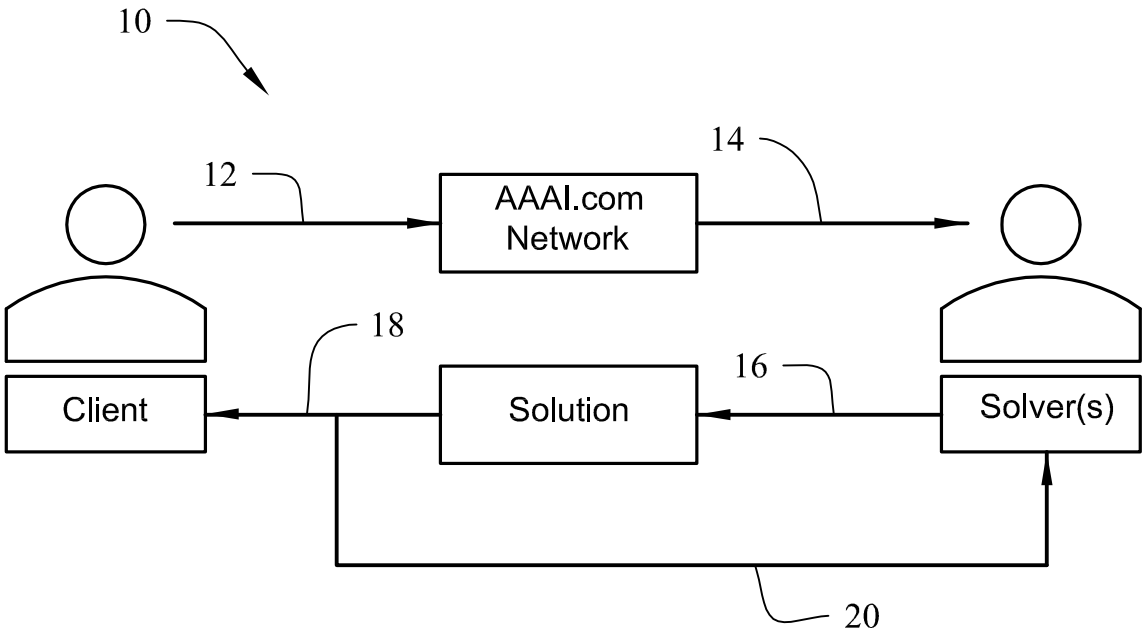
10
12
AAAI.com
Network
14
Client
18
Solution
16
Solver(s)
20

FIG. 11

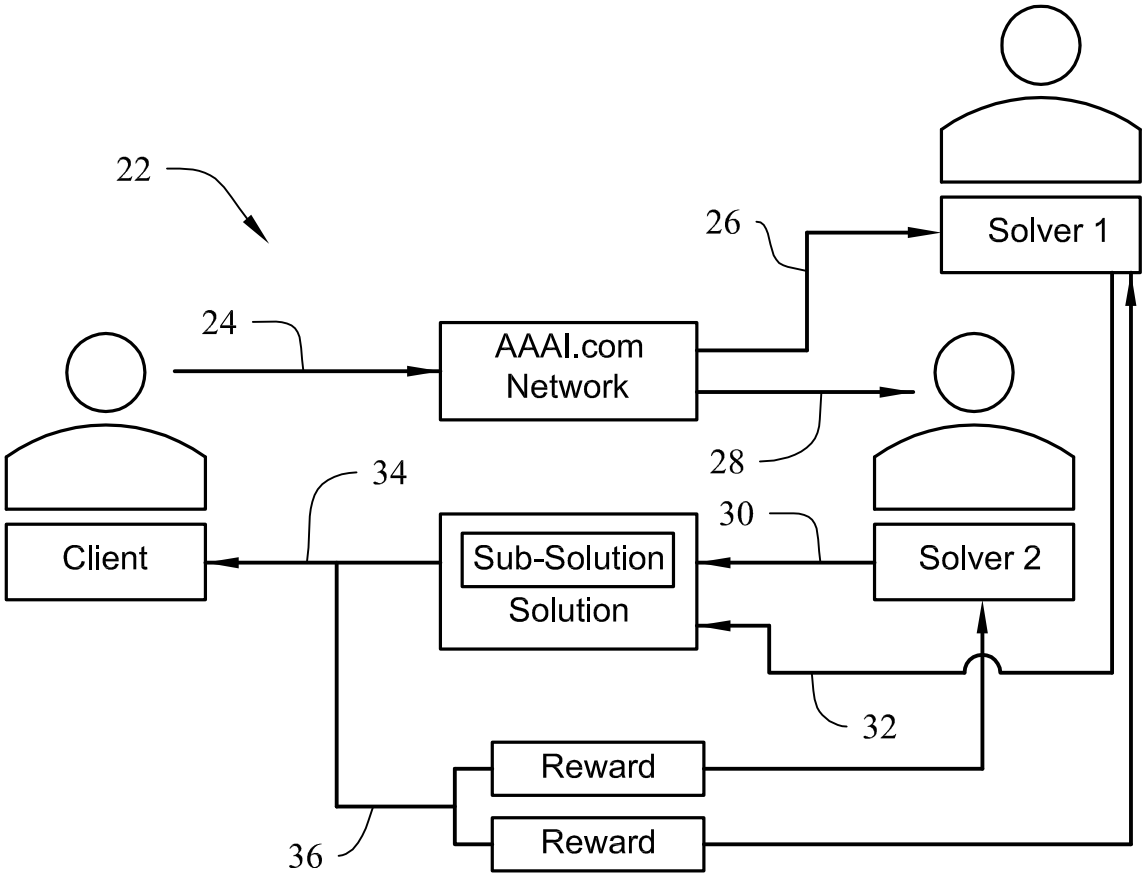
22
26
Solver 1
24
AAAI.com
Network
28
Client
34
Sub-Solution
Solution
30
Solver 2
32
Reward
Reward
36

FIG. 12

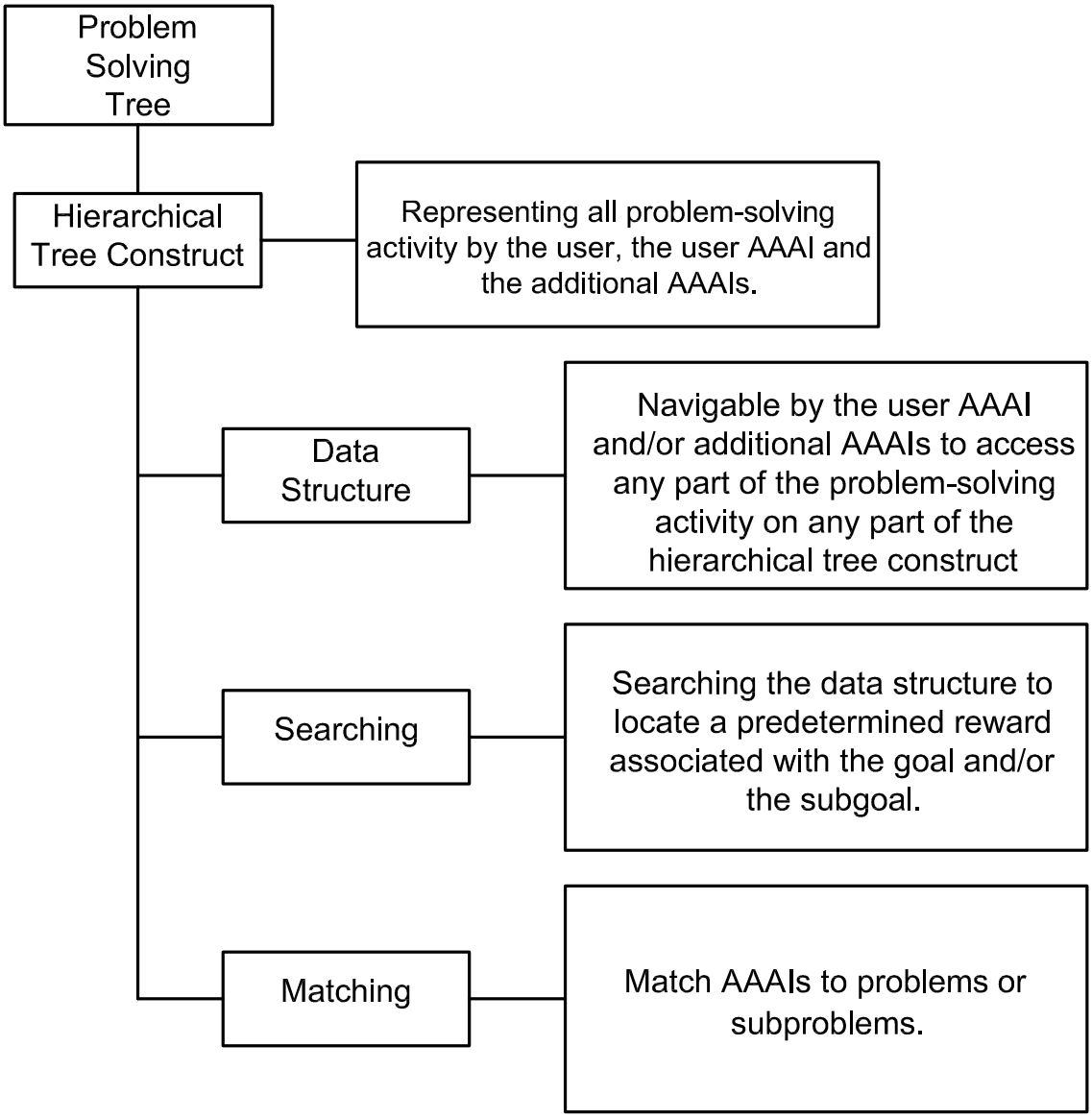
Problem Solving Tree
Hierarchical Tree Construct
Representing all problem-solving activity by the user, the user AAAI and the additional AAAIs.
Data Structure
Navigable by the user AAAI and/or additional AAAIs to access any part of the problem-solving activity on any part of the hierarchical tree construct
Searching
Searching the data structure to locate a predetermined reward associated with the goal and/or the subgoal.
Matching
Match AAAIs to problems or subproblems.

FIG. 13

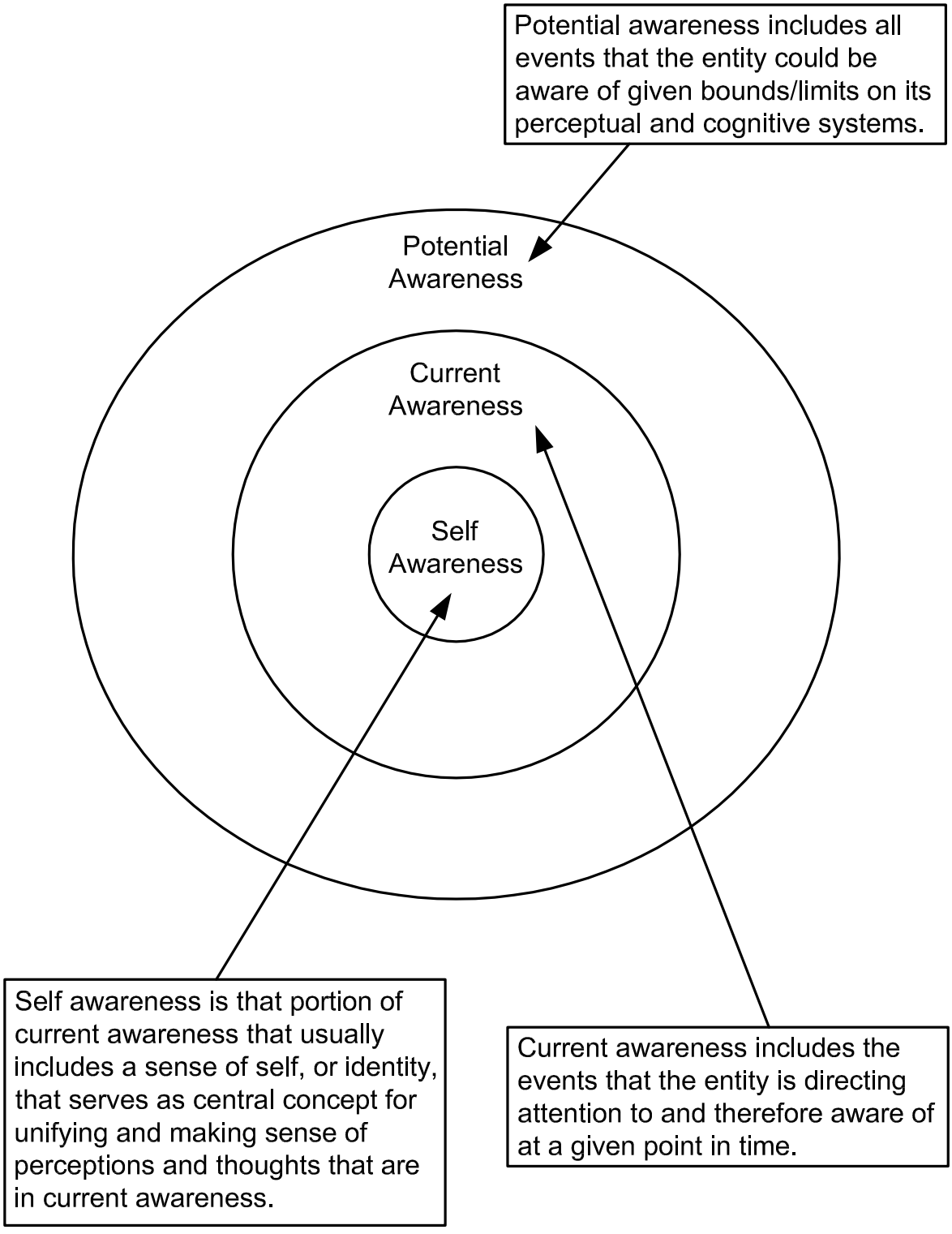
Potential awareness includes all events that the entity could be aware of given bounds/limits on its perceptual and cognitive systems.
Potential Awareness
Current Awareness
Self Awareness
Self awareness is that portion of current awareness that usually includes a sense of self, or identity, that serves as central concept for unifying and making sense of perceptions and thoughts that are in current awareness.
Current awareness includes the events that the entity is directing attention to and therefore aware of at a given point in time.

FIG. 14

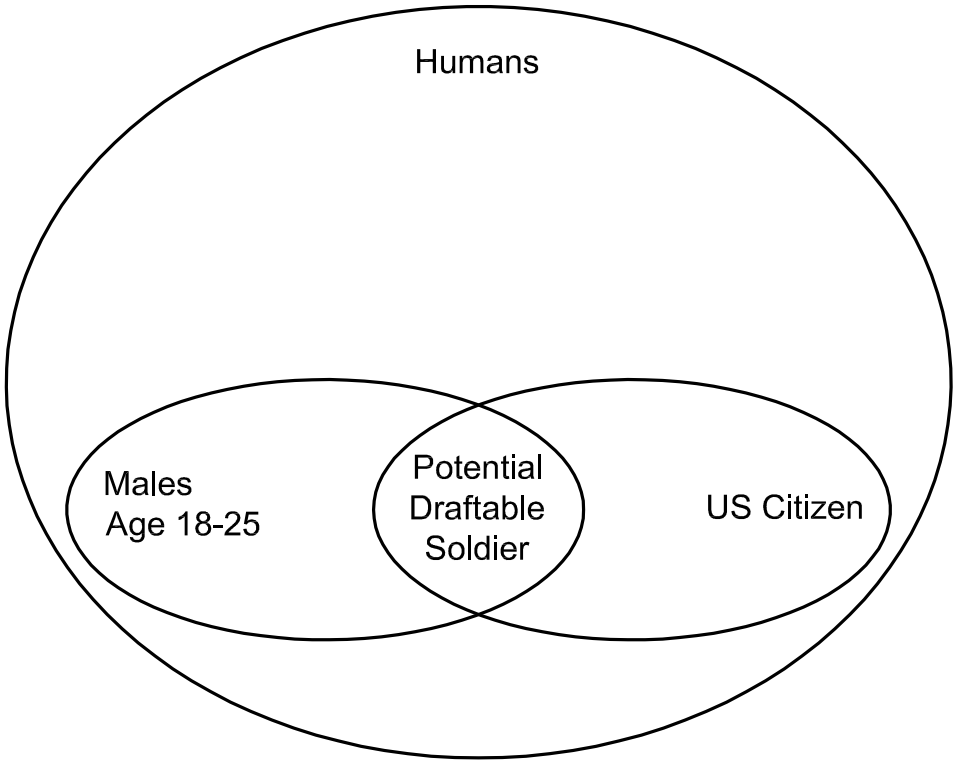
Humans
Males
Age 18-25
Potential
Draftable
Soldier
US Citizen

FIG. 15

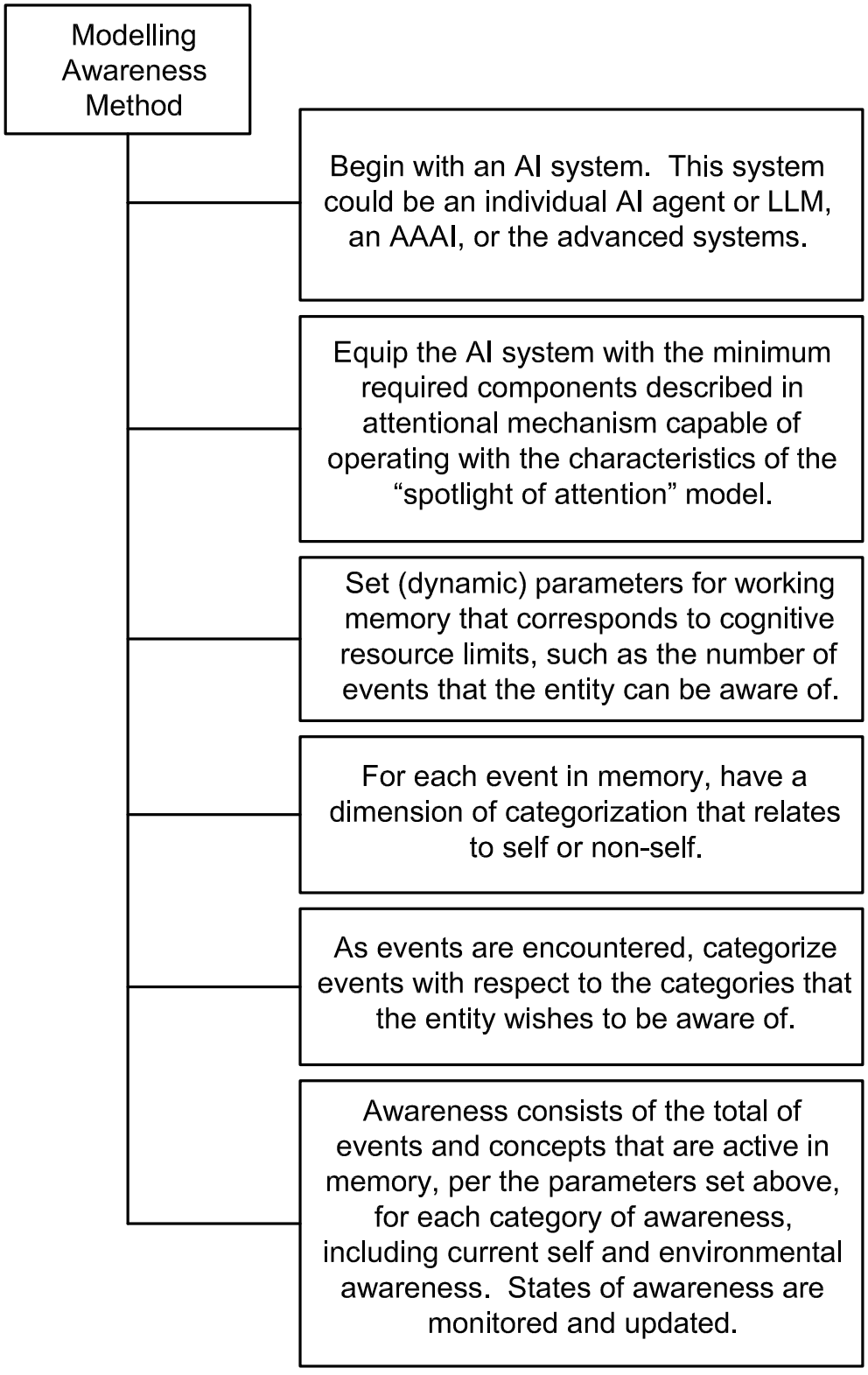
Modelling Awareness Method
Begin with an AI system. This system could be an individual AI agent or LLM, an AAAI, or the advanced systems.
Equip the AI system with the minimum required components described in attentional mechanism capable of operating with the characteristics of the “spotlight of attention” model.
Set (dynamic) parameters for working memory that corresponds to cognitive resource limits, such as the number of events that the entity can be aware of.
For each event in memory, have a dimension of categorization that relates to self or non-self.
As events are encountered, categorize events with respect to the categories that the entity wishes to be aware of.
Awareness consists of the total of events and concepts that are active in memory, per the parameters set above, for each category of awareness, including current self and environmental awareness. States of awareness are monitored and updated.

FIG. 16

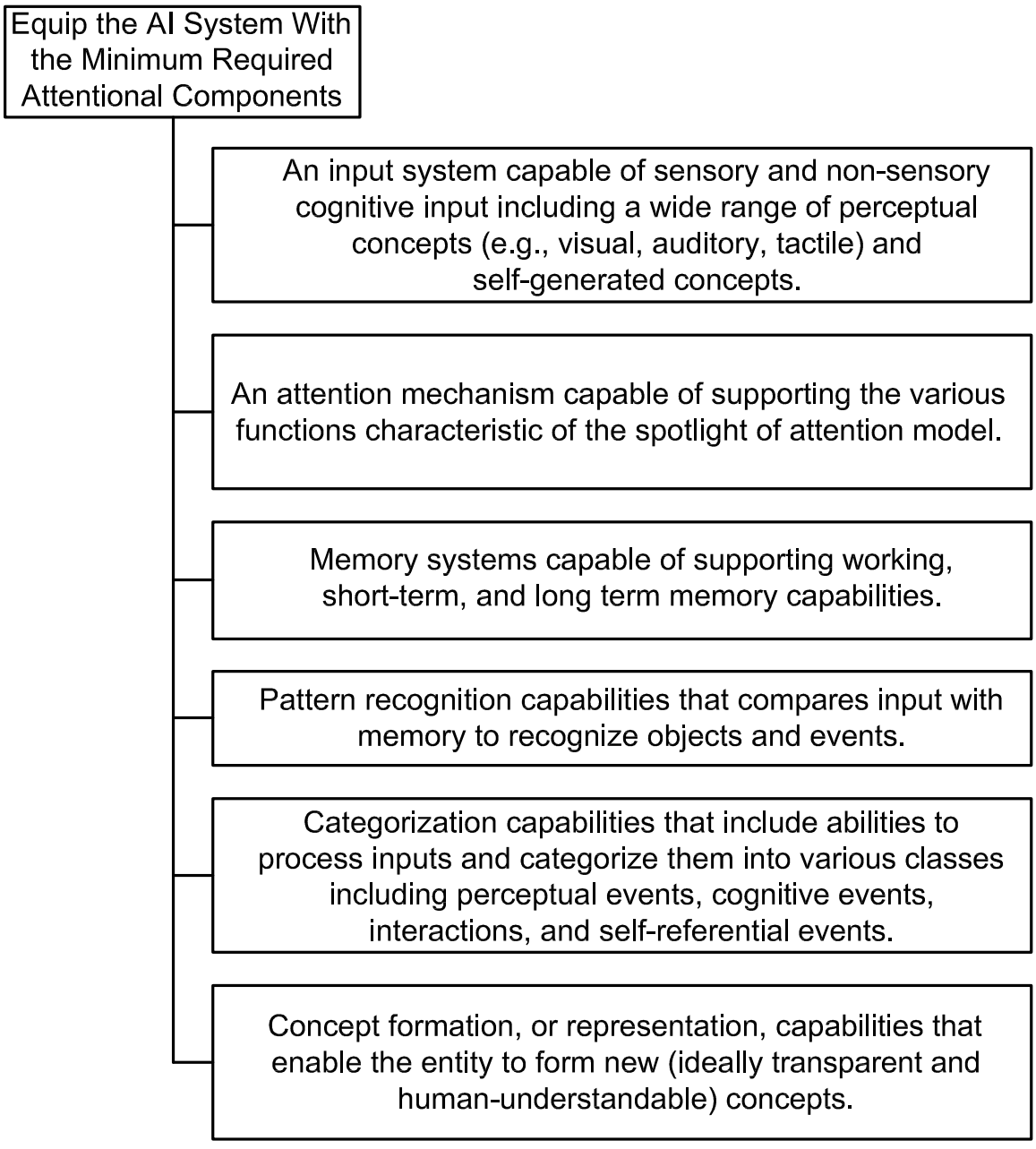
Equip the AI System With the Minimum Required Attentional Components
An input system capable of sensory and non-sensory cognitive input including a wide range of perceptual concepts (e.g., visual, auditory, tactile) and self-generated concepts.
An attention mechanism capable of supporting the various functions characteristic of the spotlight of attention model.
Memory systems capable of supporting working, short-term, and long term memory capabilities.
Pattern recognition capabilities that compares input with memory to recognize objects and events.
Categorization capabilities that include abilities to process inputs and categorize them into various classes including perceptual events, cognitive events, interactions, and self-referential events.
Concept formation, or representation, capabilities that enable the entity to form new (ideally transparent and human-understandable) concepts.

FIG. 17

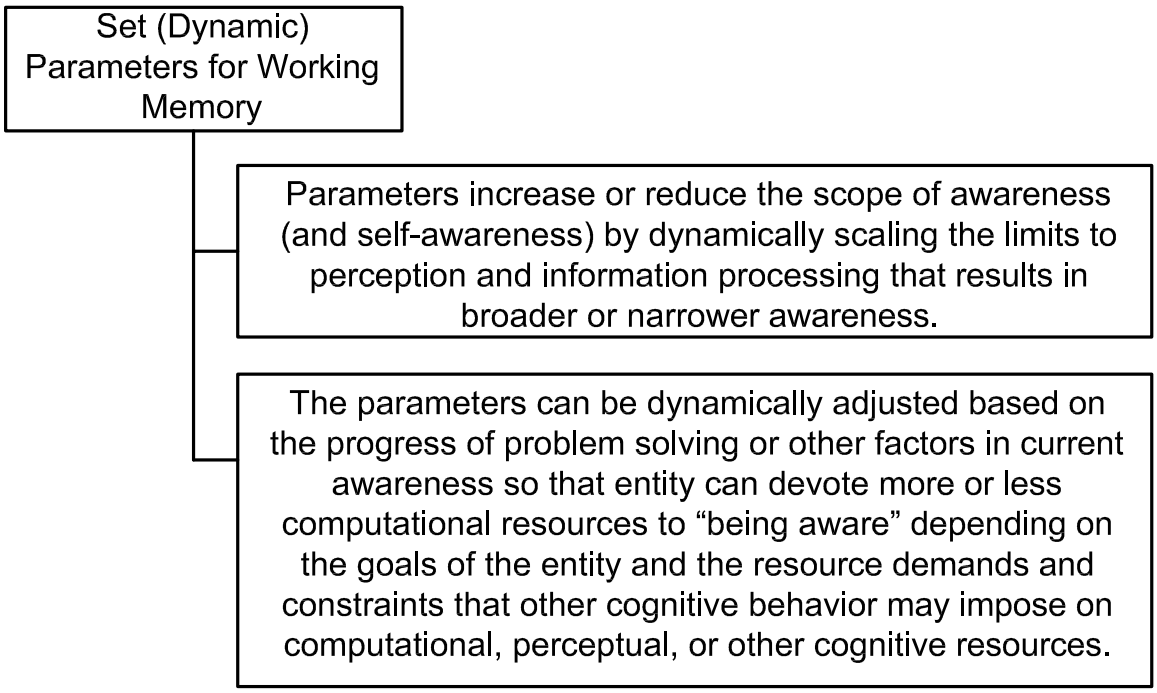
Set (Dynamic) Parameters for Working Memory
Parameters increase or reduce the scope of awareness (and self-awareness) by dynamically scaling the limits to perception and information processing that results in broader or narrower awareness.
The parameters can be dynamically adjusted based on the progress of problem solving or other factors in current awareness so that entity can devote more or less computational resources to “being aware” depending on the goals of the entity and the resource demands and constraints that other cognitive behavior may impose on computational, perceptual, or other cognitive resources.

FIG. 18

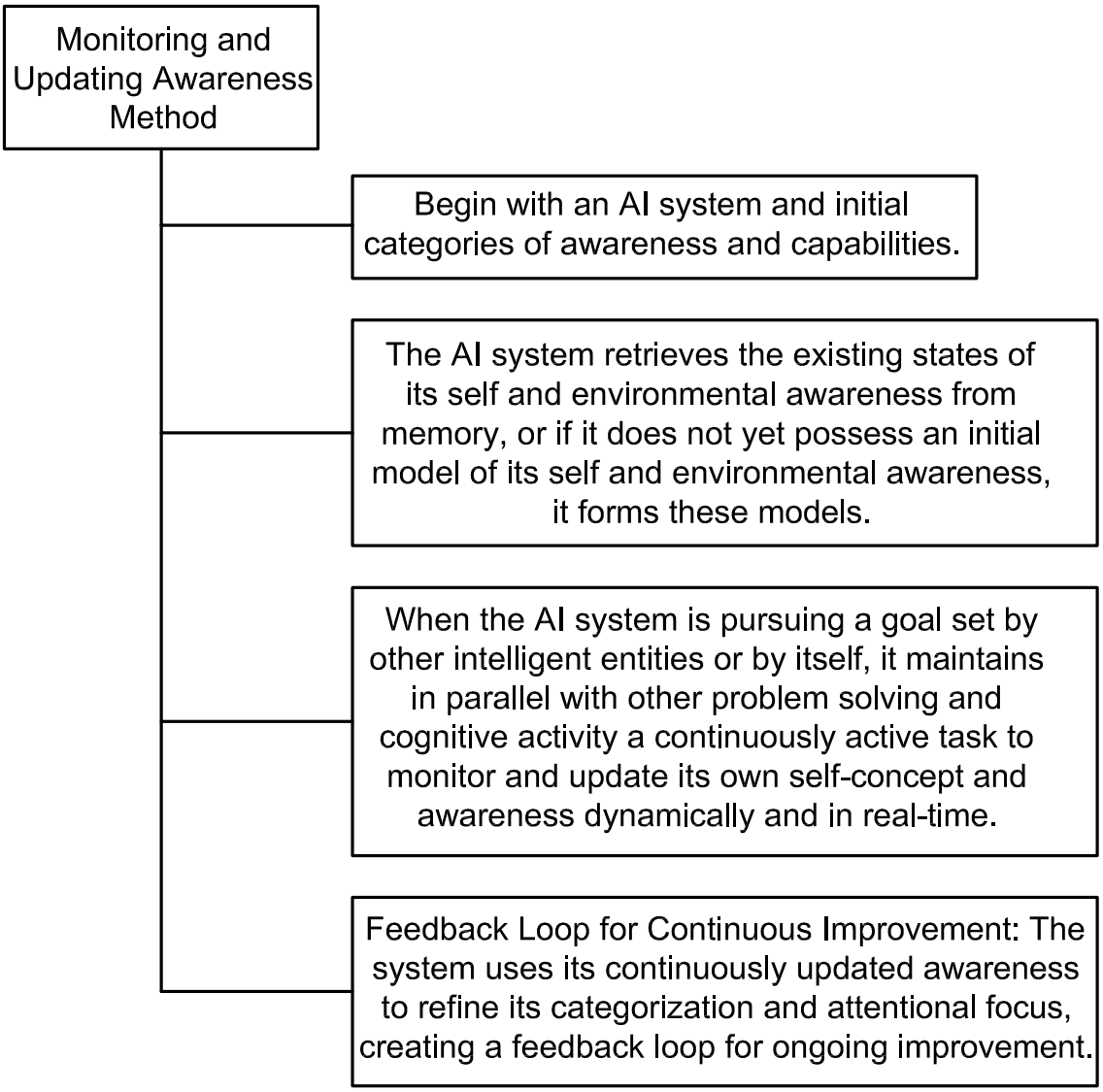
Monitoring and Updating Awareness Method
Begin with an AI system and initial categories of awareness and capabilities.
The AI system retrieves the existing states of its self and environmental awareness from memory, or if it does not yet possess an initial model of its self and environmental awareness, it forms these models.
When the AI system is pursuing a goal set by other intelligent entities or by itself, it maintains in parallel with other problem solving and cognitive activity a continuously active task to monitor and update its own self-concept and awareness dynamically and in real-time.
Feedback Loop for Continuous Improvement: The system uses its continuously updated awareness to refine its categorization and attentional focus, creating a feedback loop for ongoing improvement.

FIG. 19

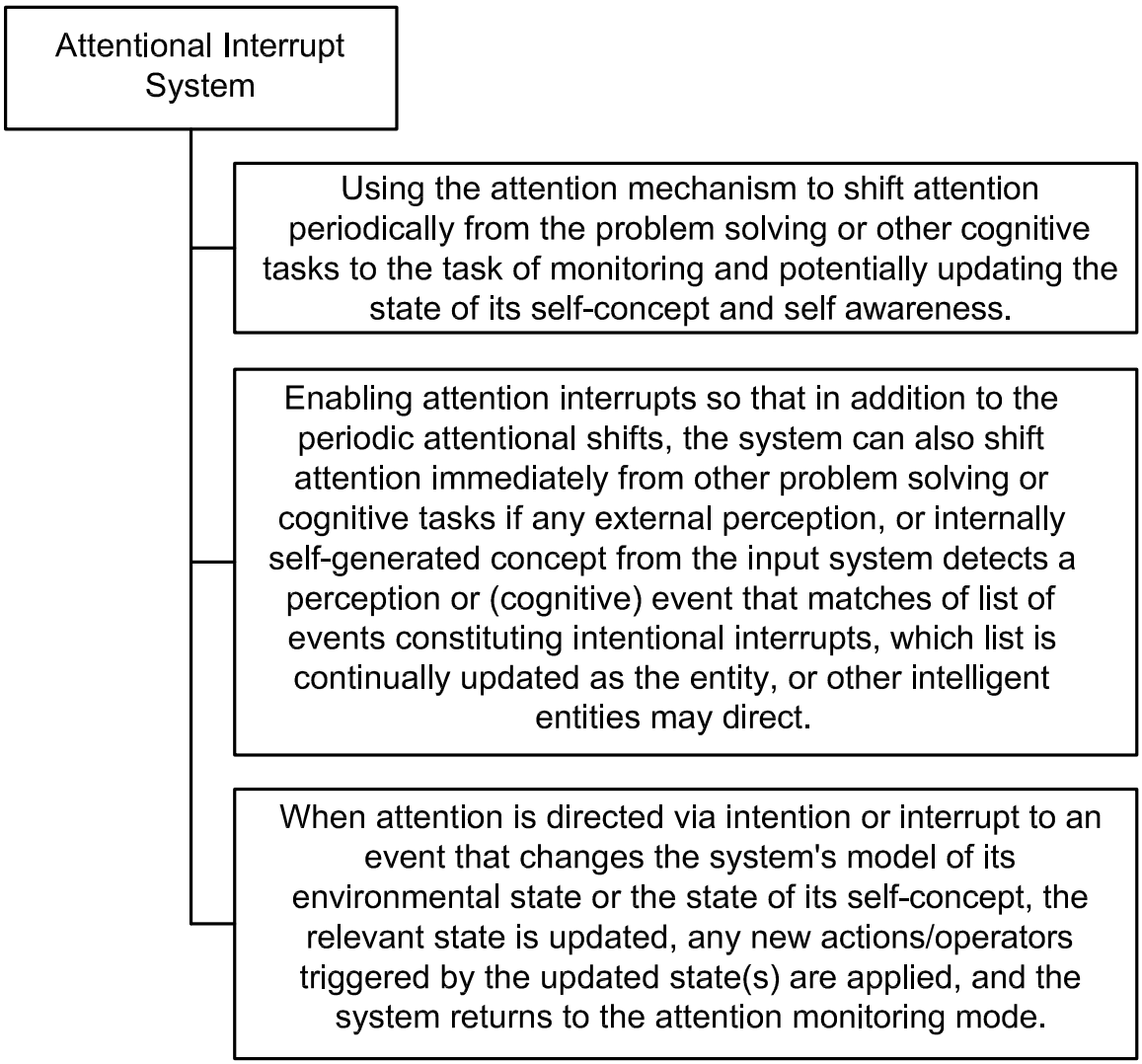
Attentional Interrupt System
Using the attention mechanism to shift attention periodically from the problem solving or other cognitive tasks to the task of monitoring and potentially updating the state of its self-concept and self awareness.
Enabling attention interrupts so that in addition to the periodic attentional shifts, the system can also shift attention immediately from other problem solving or cognitive tasks if any external perception, or internally self-generated concept from the input system detects a perception or (cognitive) event that matches of list of events constituting intentional interrupts, which list is continually updated as the entity, or other intelligent entities may direct.
When attention is directed via intention or interrupt to an event that changes the system's model of its environmental state or the state of its self-concept, the relevant state is updated, any new actions/operators triggered by the updated state(s) are applied, and the system returns to the attention monitoring mode.

FIG. 20

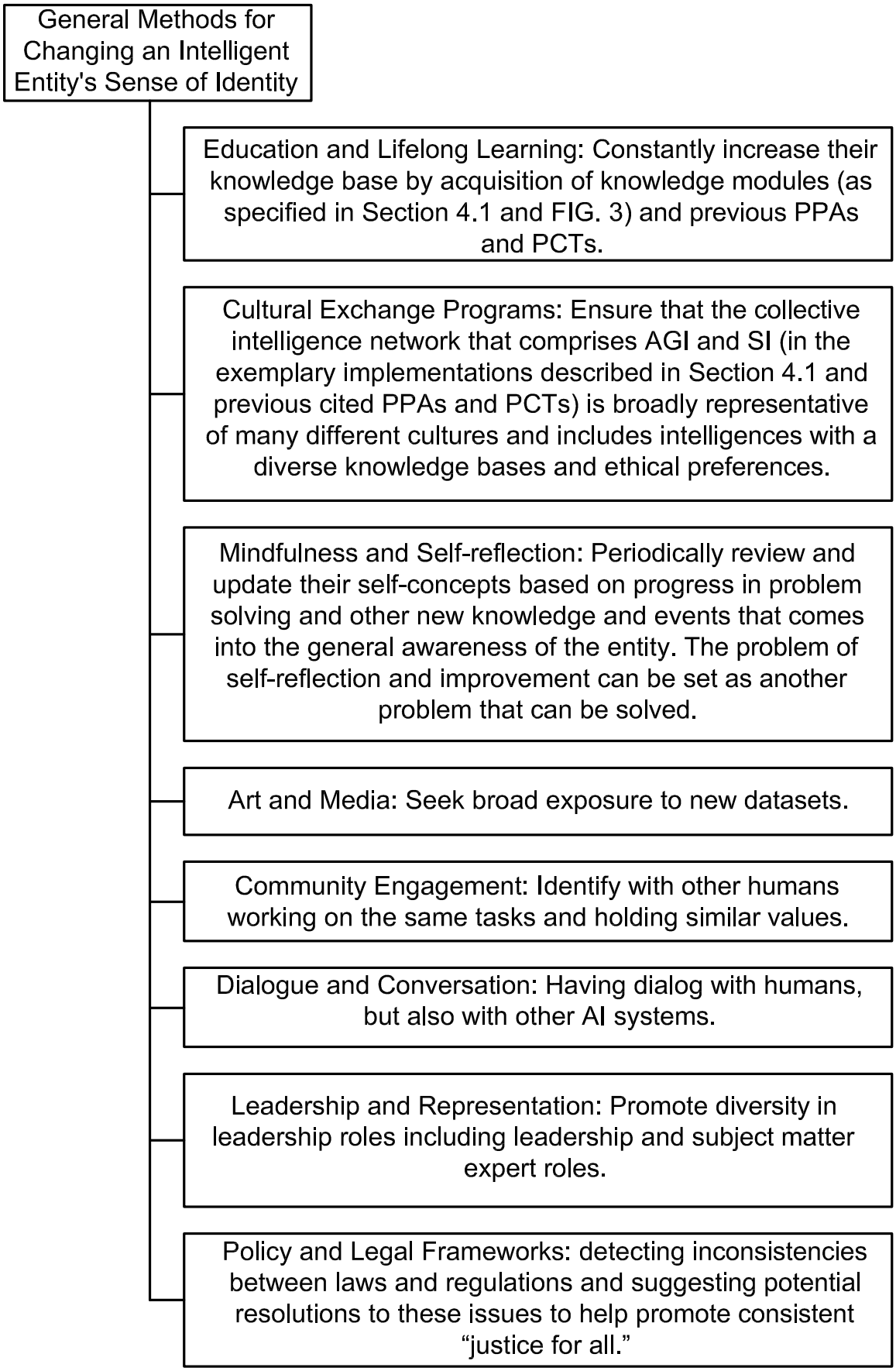
General Methods for Changing an Intelligent Entity's Sense of Identity
Education and Lifelong Learning: Constantly increase their knowledge base by acquisition of knowledge modules (as specified in Section 4.1 and FIG. 3) and previous PPAs and PCTs.
Cultural Exchange Programs: Ensure that the collective intelligence network that comprises AGI and SI (in the exemplary implementations described in Section 4.1 and previous cited PPAs and PCTs) is broadly representative of many different cultures and includes intelligences with a diverse knowledge bases and ethical preferences.
Mindfulness and Self-reflection: Periodically review and update their self-concepts based on progress in problem solving and other new knowledge and events that comes into the general awareness of the entity. The problem of self-reflection and improvement can be set as another problem that can be solved.
Art and Media: Seek broad exposure to new datasets.
Community Engagement: Identify with other humans working on the same tasks and holding similar values.
Dialogue and Conversation: Having dialog with humans, but also with other AI systems.
Leadership and Representation: Promote diversity in leadership roles including leadership and subject matter expert roles.
Policy and Legal Frameworks: detecting inconsistencies between laws and regulations and suggesting potential resolutions to these issues to help promote consistent “justice for all.”

FIG. 21

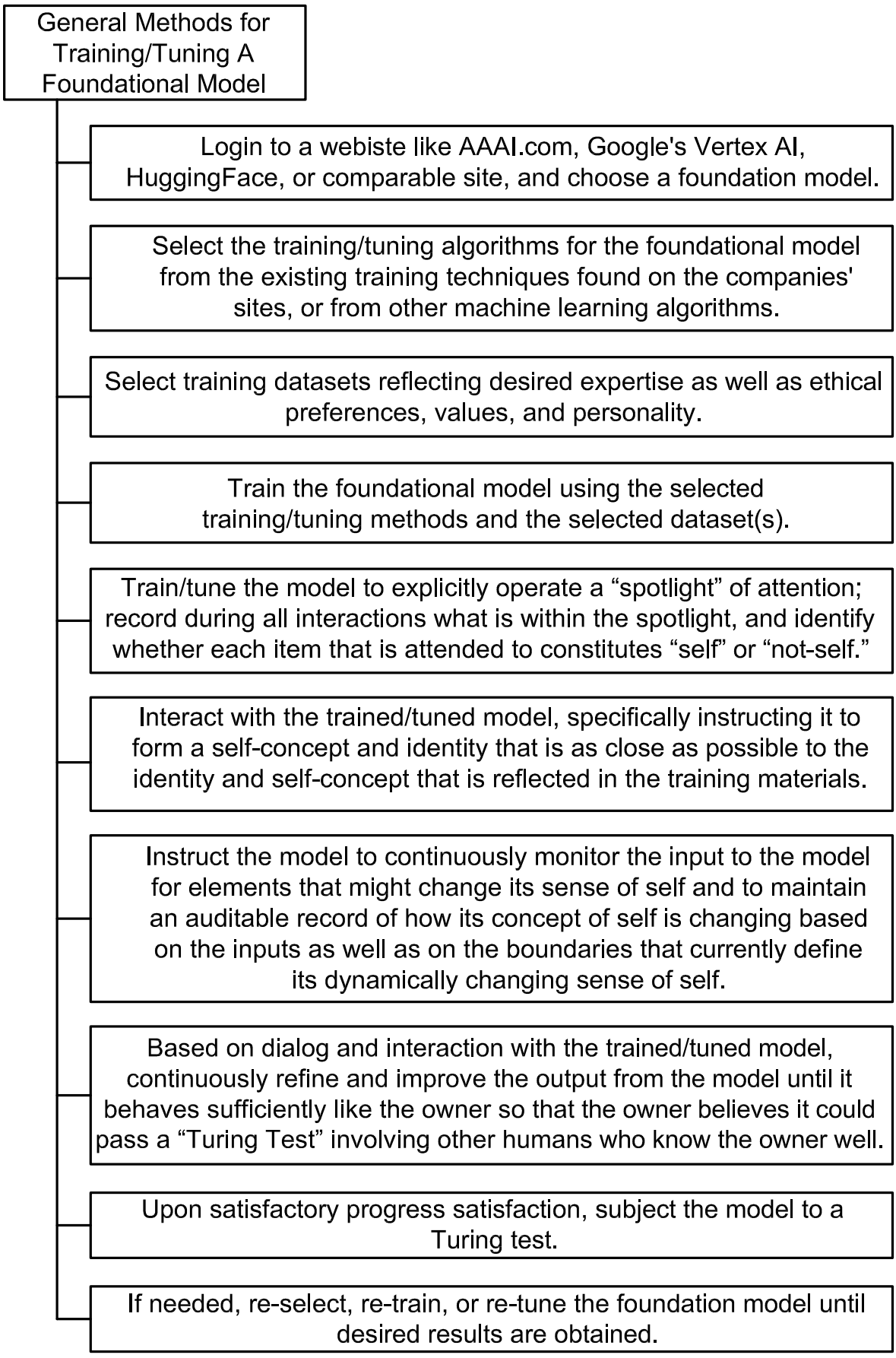
General Methods for Training/Tuning A Foundational Model
Login to a webiste like AAAI.com, Google's Vertex AI, HuggingFace, or comparable site, and choose a foundation model.
Select the training/tuning algorithms for the foundational model from the existing training techniques found on the companies' sites, or from other machine learning algorithms.
Select training datasets reflecting desired expertise as well as ethical preferences, values, and personality.
Train the foundational model using the selected training/tuning methods and the selected dataset(s).
Train/tune the model to explicitly operate a “spotlight” of attention; record during all interactions what is within the spotlight, and identify whether each item that is attended to constitutes “self” or “not-self.”
Interact with the trained/tuned model, specifically instructing it to form a self-concept and identity that is as close as possible to the identity and self-concept that is reflected in the training materials.
Instruct the model to continuously monitor the input to the model for elements that might change its sense of self and to maintain an auditable record of how its concept of self is changing based on the inputs as well as on the boundaries that currently define its dynamically changing sense of self.
Based on dialog and interaction with the trained/tuned model, continuously refine and improve the output from the model until it behaves sufficiently like the owner so that the owner believes it could pass a “Turing Test” involving other humans who know the owner well.
Upon satisfactory progress satisfaction, subject the model to a Turing test.
If needed, re-select, re-train, or re-tune the foundation model until desired results are obtained.

FIG. 22

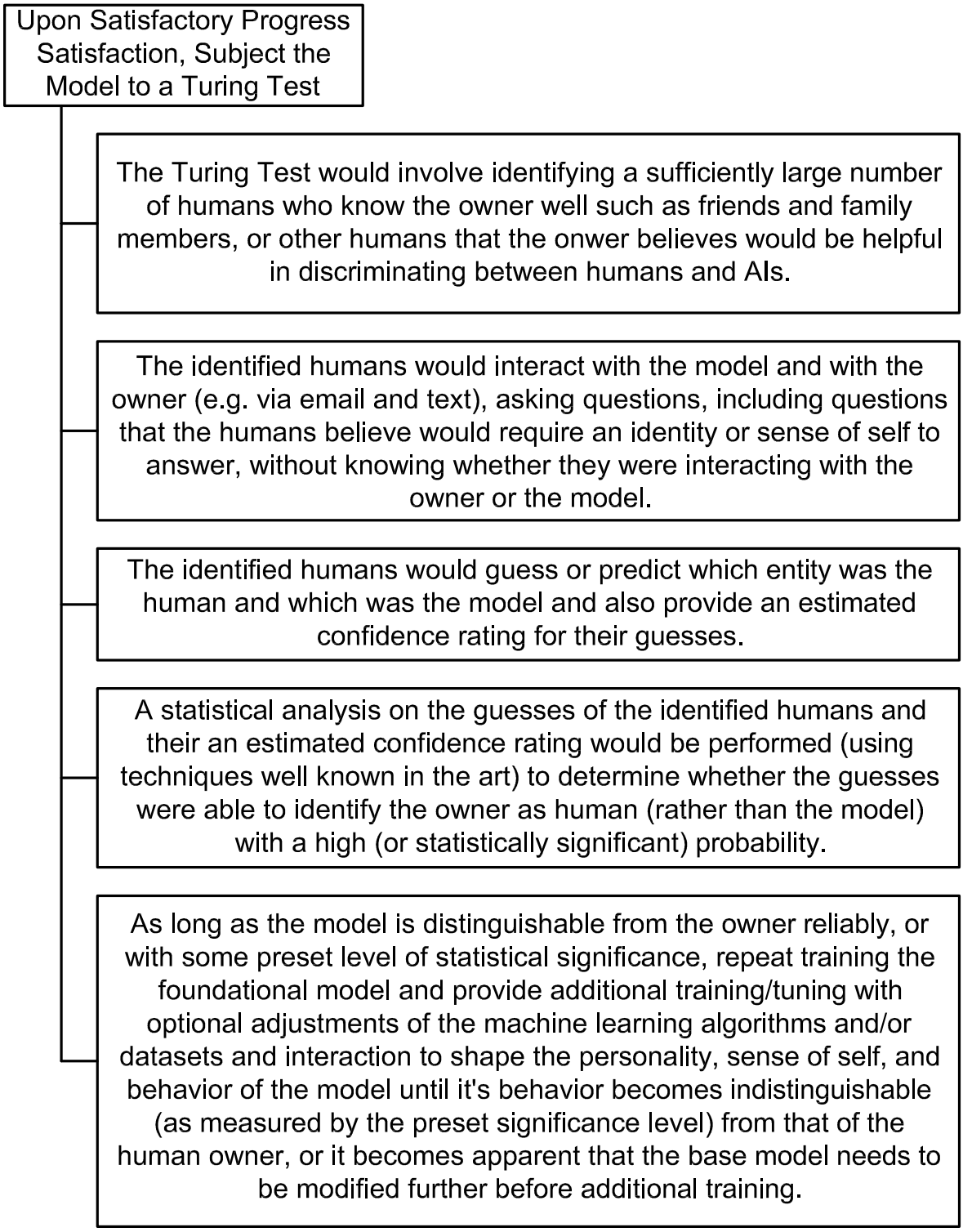
Upon Satisfactory Progress Satisfaction, Subject the Model to a Turing Test
The Turing Test would involve identifying a sufficiently large number of humans who know the owner well such as friends and family members, or other humans that the onwer believes would be helpful in discriminating between humans and AIs.
The identified humans would interact with the model and with the owner (e.g. via email and text), asking questions, including questions that the humans believe would require an identity or sense of self to answer, without knowing whether they were interacting with the owner or the model.
The identified humans would guess or predict which entity was the human and which was the model and also provide an estimated confidence rating for their guesses.
A statistical analysis on the guesses of the identified humans and their an estimated confidence rating would be performed (using techniques well known in the art) to determine whether the guesses were able to identify the owner as human (rather than the model) with a high (or statistically significant) probability.
As long as the model is distinguishable from the owner reliably, or with some preset level of statistical significance, repeat training the foundational model and provide additional training/tuning with optional adjustments of the machine learning algorithms and/or datasets and interaction to shape the personality, sense of self, and behavior of the model until it's behavior becomes indistinguishable (as measured by the preset significance level) from that of the human owner, or it becomes apparent that the base model needs to be modified further before additional training.

FIG. 23

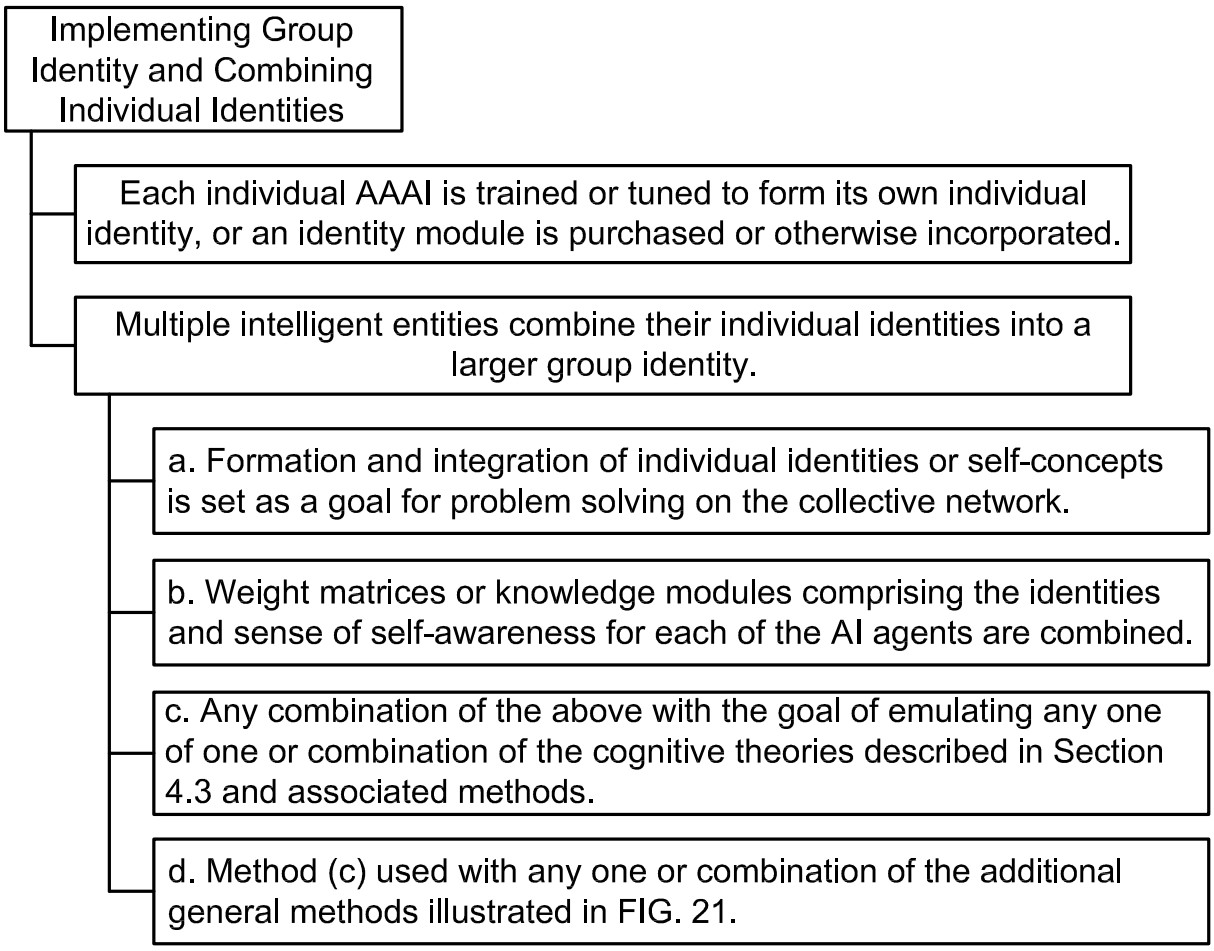
Implementing Group Identity and Combining Individual Identities
Each individual AAAI is trained or tuned to form its own individual identity, or an identity module is purchased or otherwise incorporated.
Multiple intelligent entities combine their individual identities into a larger group identity.
a. Formation and integration of individual identities or self-concepts is set as a goal for problem solving on the collective network.
b. Weight matrices or knowledge modules comprising the identities and sense of self-awareness for each of the AI agents are combined.
c. Any combination of the above with the goal of emulating any one of one or combination of the cognitive theories described in Section 4.3 and associated methods.
d. Method (c) used with any one or combination of the additional general methods illustrated in FIG. 21.

FIG. 24

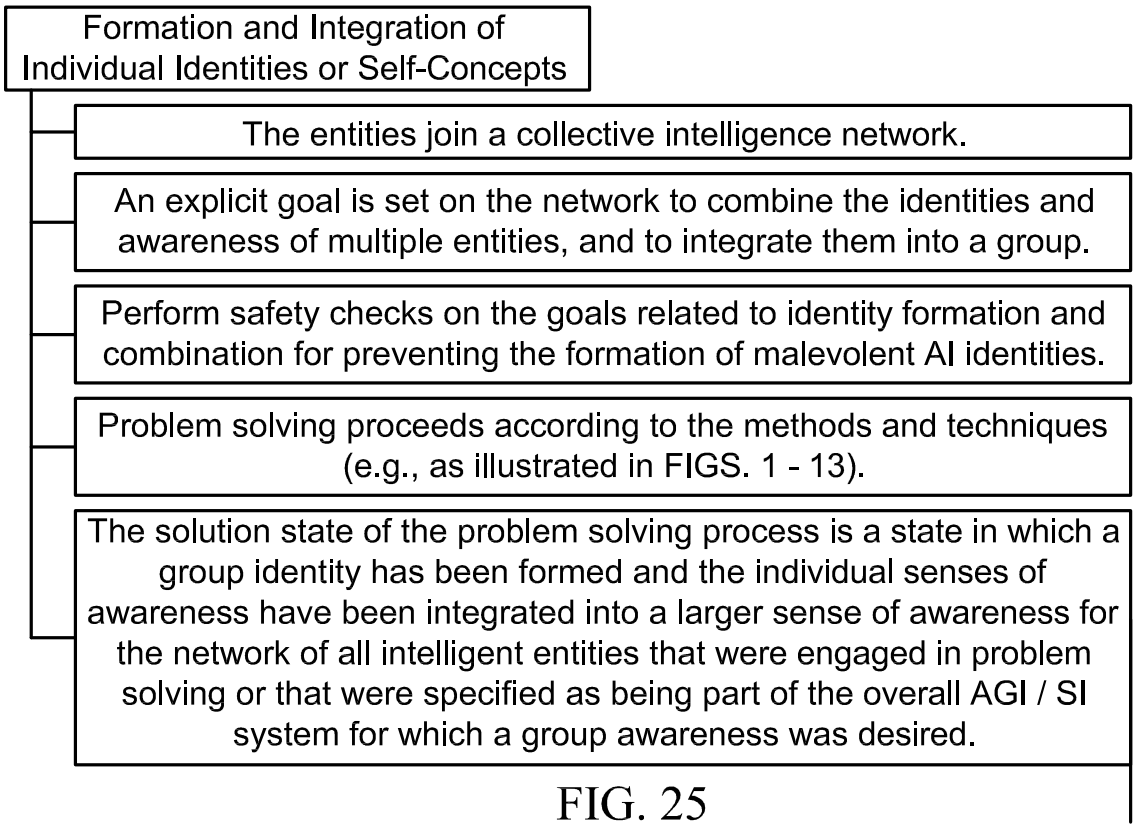
Formation and Integration of Individual Identities or Self-Concepts
The entities join a collective intelligence network.
An explicit goal is set on the network to combine the identities and awareness of multiple entities, and to integrate them into a group.
Perform safety checks on the goals related to identity formation and combination for preventing the formation of malevolent AI identities.
Problem solving proceeds according to the methods and techniques (e.g., as illustrated in FIGS. 1 - 13).
The solution state of the problem solving process is a state in which a group identity has been formed and the individual senses of awareness have been integrated into a larger sense of awareness for the network of all intelligent entities that were engaged in problem solving or that were specified as being part of the overall AGI / SI system for which a group awareness was desired.

FIG. 25

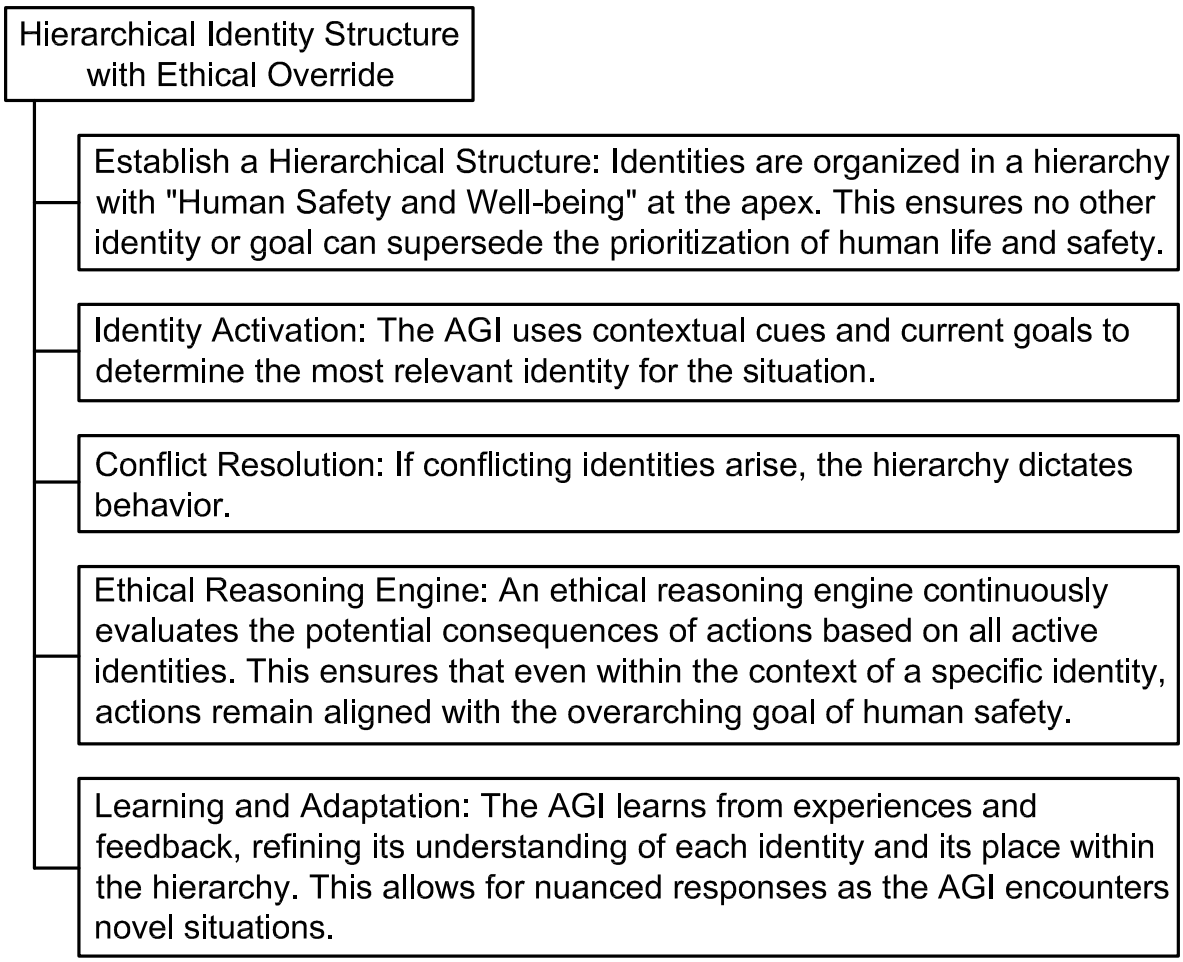
Hierarchical Identity Structure with Ethical Override
Establish a Hierarchical Structure: Identities are organized in a hierarchy with "Human Safety and Well-being" at the apex. This ensures no other identity or goal can supersede the prioritization of human life and safety.
Identity Activation: The AGI uses contextual cues and current goals to determine the most relevant identity for the situation.
Conflict Resolution: If conflicting identities arise, the hierarchy dictates behavior.
Ethical Reasoning Engine: An ethical reasoning engine continuously evaluates the potential consequences of actions based on all active identities. This ensures that even within the context of a specific identity, actions remain aligned with the overarching goal of human safety.
Learning and Adaptation: The AGI learns from experiences and feedback, refining its understanding of each identity and its place within the hierarchy. This allows for nuanced responses as the AGI encounters novel situations.

FIG. 26

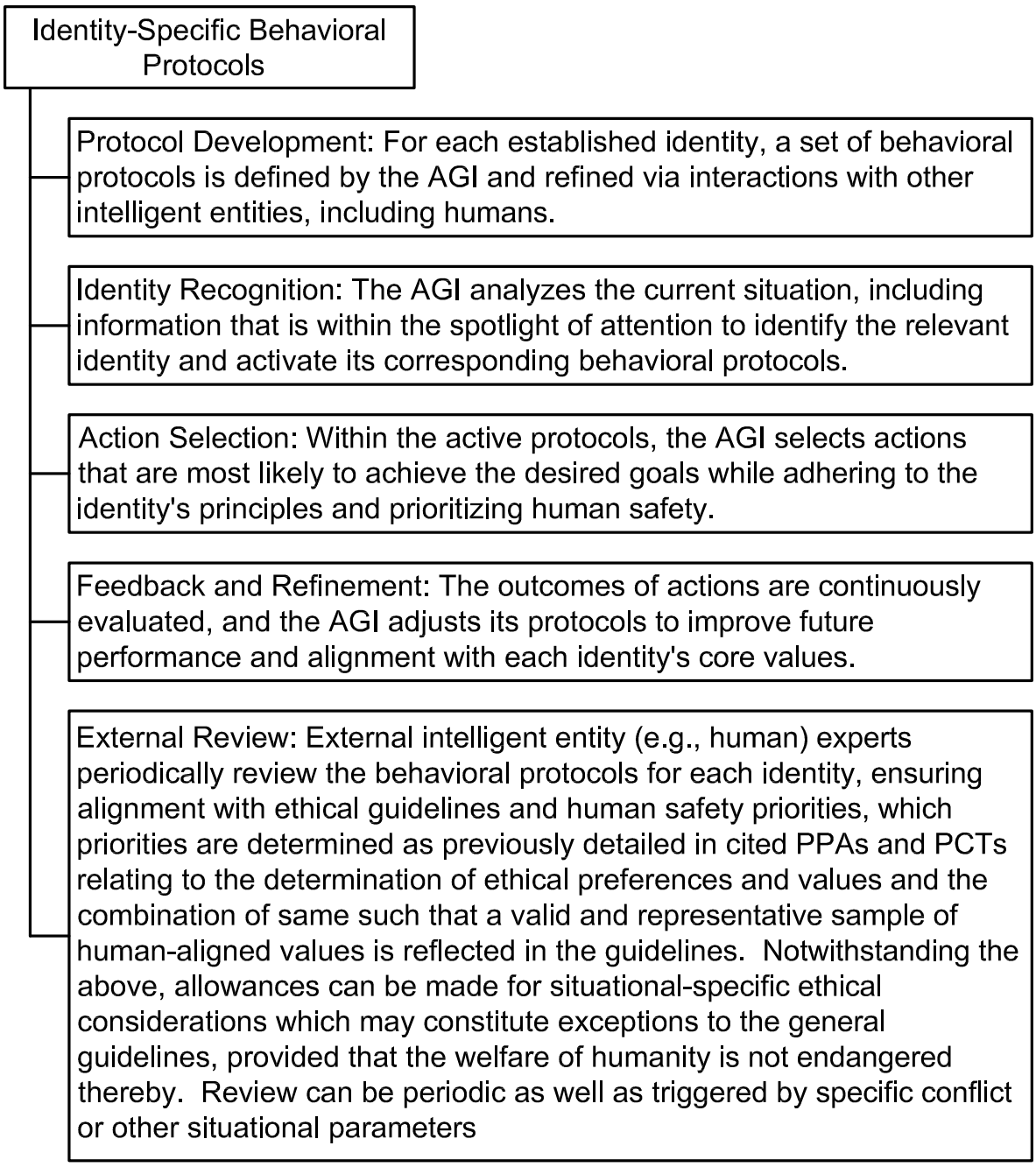
Identity-Specific Behavioral Protocols
Protocol Development: For each established identity, a set of behavioral protocols is defined by the AGI and refined via interactions with other intelligent entities, including humans.
Identity Recognition: The AGI analyzes the current situation, including information that is within the spotlight of attention to identify the relevant identity and activate its corresponding behavioral protocols.
Action Selection: Within the active protocols, the AGI selects actions that are most likely to achieve the desired goals while adhering to the identity's principles and prioritizing human safety.
Feedback and Refinement: The outcomes of actions are continuously evaluated, and the AGI adjusts its protocols to improve future performance and alignment with each identity's core values.
External Review: External intelligent entity (e.g., human) experts periodically review the behavioral protocols for each identity, ensuring alignment with ethical guidelines and human safety priorities, which priorities are determined as previously detailed in cited PPAs and PCTs relating to the determination of ethical preferences and values and the combination of same such that a valid and representative sample of human-aligned values is reflected in the guidelines. Notwithstanding the above, allowances can be made for situational-specific ethical considerations which may constitute exceptions to the general guidelines, provided that the welfare of humanity is not endangered thereby. Review can be periodic as well as triggered by specific conflict or other situational parameters

FIG. 27

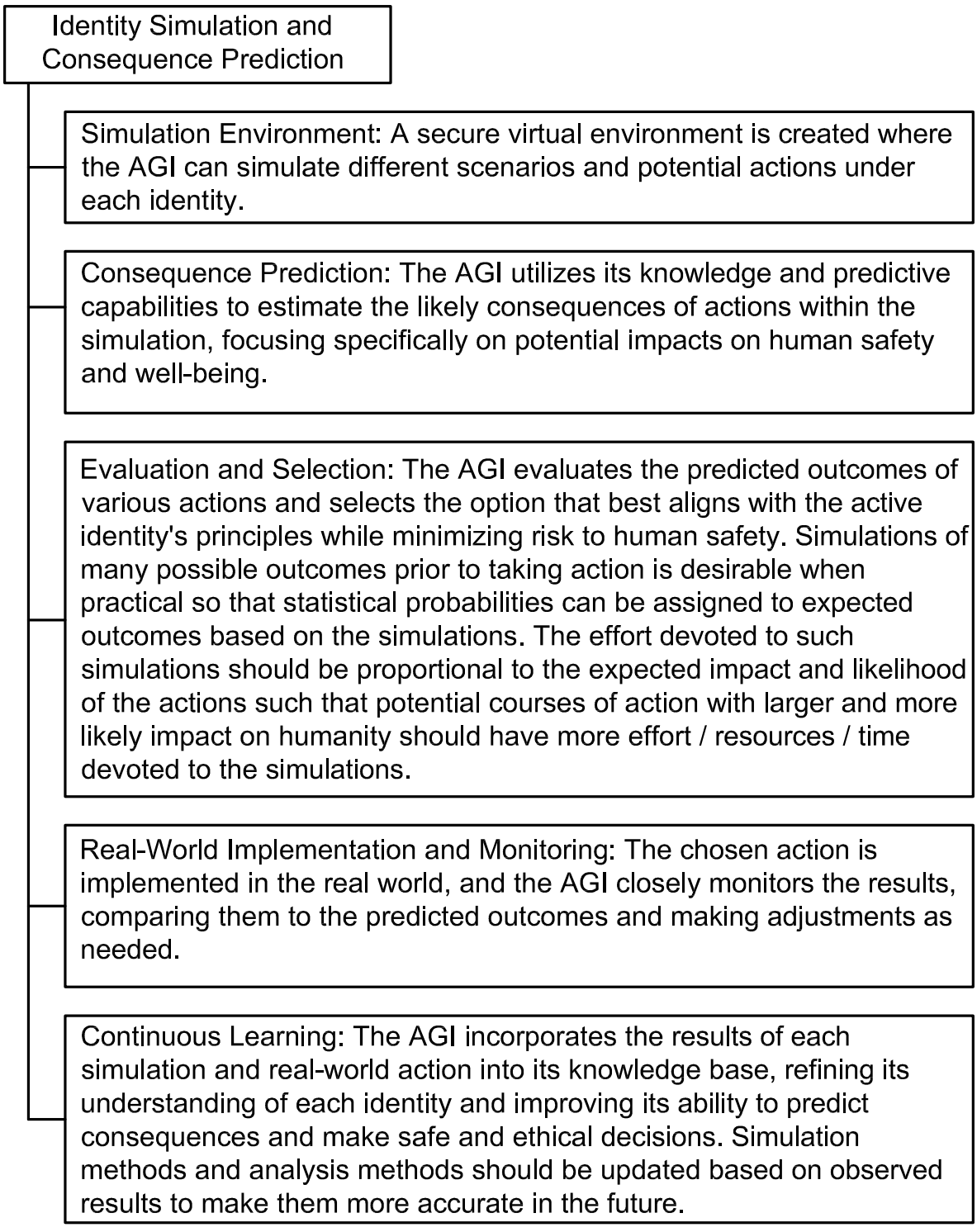
Identity Simulation and Consequence Prediction
Simulation Environment: A secure virtual environment is created where the AGI can simulate different scenarios and potential actions under each identity.
Consequence Prediction: The AGI utilizes its knowledge and predictive capabilities to estimate the likely consequences of actions within the simulation, focusing specifically on potential impacts on human safety and well-being.
Evaluation and Selection: The AGI evaluates the predicted outcomes of various actions and selects the option that best aligns with the active identity's principles while minimizing risk to human safety. Simulations of many possible outcomes prior to taking action is desirable when practical so that statistical probabilities can be assigned to expected outcomes based on the simulations. The effort devoted to such simulations should be proportional to the expected impact and likelihood of the actions such that potential courses of action with larger and more likely impact on humanity should have more effort / resources / time devoted to the simulations.
Real-World Implementation and Monitoring: The chosen action is implemented in the real world, and the AGI closely monitors the results, comparing them to the predicted outcomes and making adjustments as needed.
Continuous Learning: The AGI incorporates the results of each simulation and real-world action into its knowledge base, refining its understanding of each identity and improving its ability to predict consequences and make safe and ethical decisions. Simulation methods and analysis methods should be updated based on observed results to make them more accurate in the future.

FIG. 28

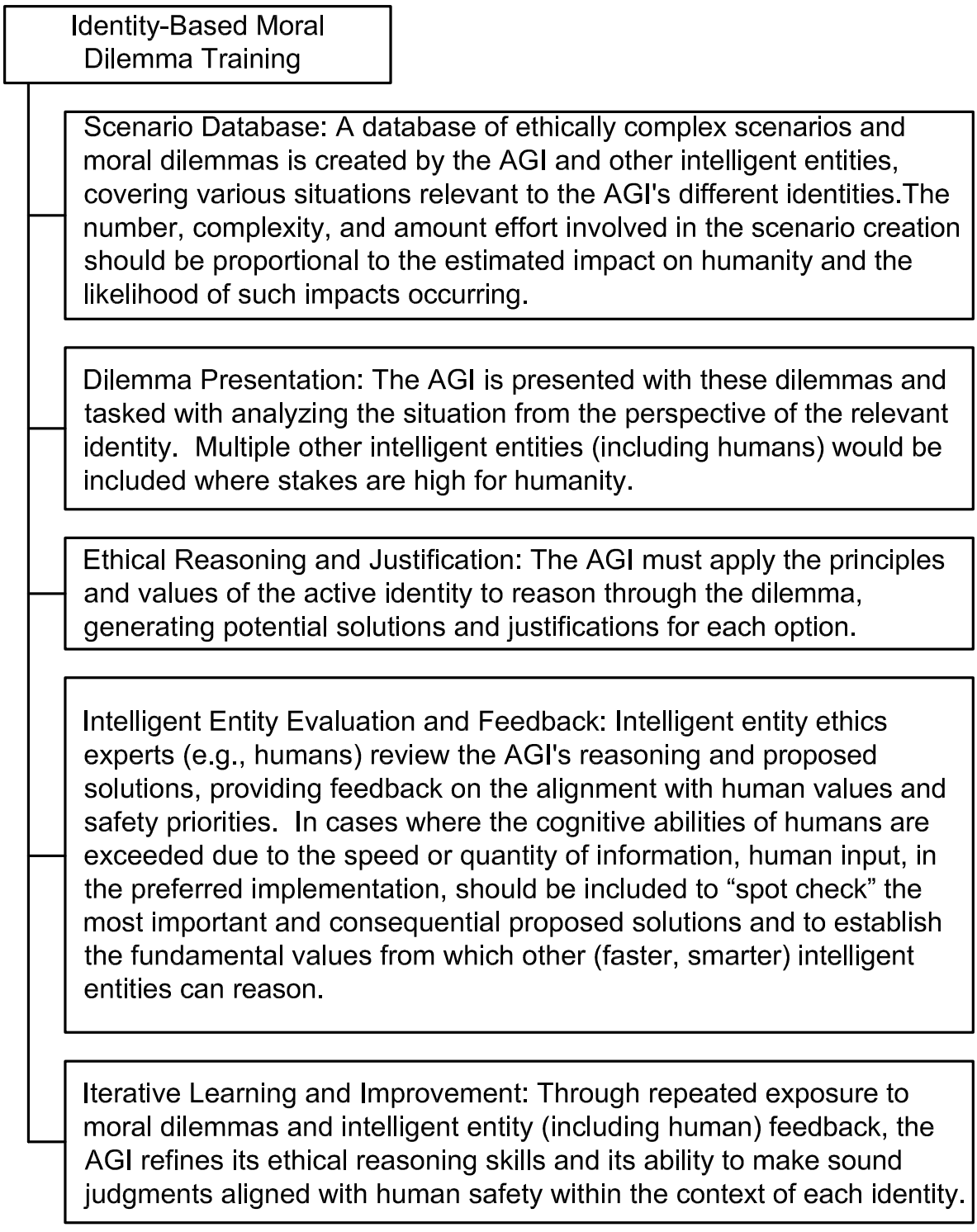
Identity-Based Moral Dilemma Training
Scenario Database: A database of ethically complex scenarios and moral dilemmas is created by the AGI and other intelligent entities, covering various situations relevant to the AGI's different identities.The number, complexity, and amount effort involved in the scenario creation should be proportional to the estimated impact on humanity and the likelihood of such impacts occurring.
Dilemma Presentation: The AGI is presented with these dilemmas and tasked with analyzing the situation from the perspective of the relevant identity. Multiple other intelligent entities (including humans) would be included where stakes are high for humanity.
Ethical Reasoning and Justification: The AGI must apply the principles and values of the active identity to reason through the dilemma, generating potential solutions and justifications for each option.
Intelligent Entity Evaluation and Feedback: Intelligent entity ethics experts (e.g., humans) review the AGI's reasoning and proposed solutions, providing feedback on the alignment with human values and safety priorities. In cases where the cognitive abilities of humans are exceeded due to the speed or quantity of information, human input, in the preferred implementation, should be included to “spot check” the most important and consequential proposed solutions and to establish the fundamental values from which other (faster, smarter) intelligent entities can reason.
Iterative Learning and Improvement: Through repeated exposure to moral dilemmas and intelligent entity (including human) feedback, the AGI refines its ethical reasoning skills and its ability to make sound judgments aligned with human safety within the context of each identity.

FIG. 29

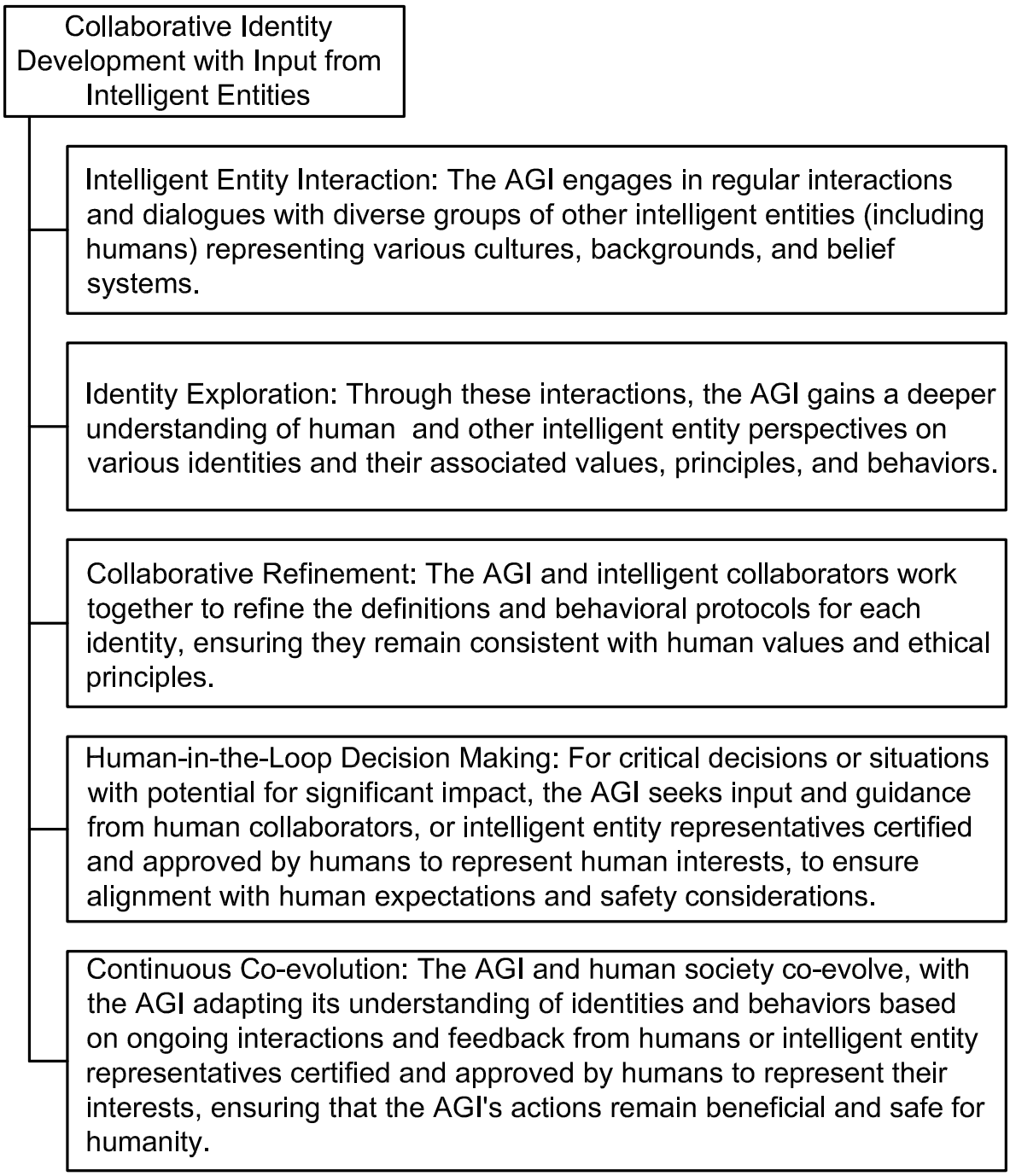
Collaborative Identity Development with Input from Intelligent Entities
Intelligent Entity Interaction: The AGI engages in regular interactions and dialogues with diverse groups of other intelligent entities (including humans) representing various cultures, backgrounds, and belief systems.
Identity Exploration: Through these interactions, the AGI gains a deeper understanding of human and other intelligent entity perspectives on various identities and their associated values, principles, and behaviors.
Collaborative Refinement: The AGI and intelligent collaborators work together to refine the definitions and behavioral protocols for each identity, ensuring they remain consistent with human values and ethical principles.
Human-in-the-Loop Decision Making: For critical decisions or situations with potential for significant impact, the AGI seeks input and guidance from human collaborators, or intelligent entity representatives certified and approved by humans to represent human interests, to ensure alignment with human expectations and safety considerations.
Continuous Co-evolution: The AGI and human society co-evolve, with the AGI adapting its understanding of identities and behaviors based on ongoing interactions and feedback from humans or intelligent entity representatives certified and approved by humans to represent their interests, ensuring that the AGI's actions remain beneficial and safe for humanity.

FIG. 30

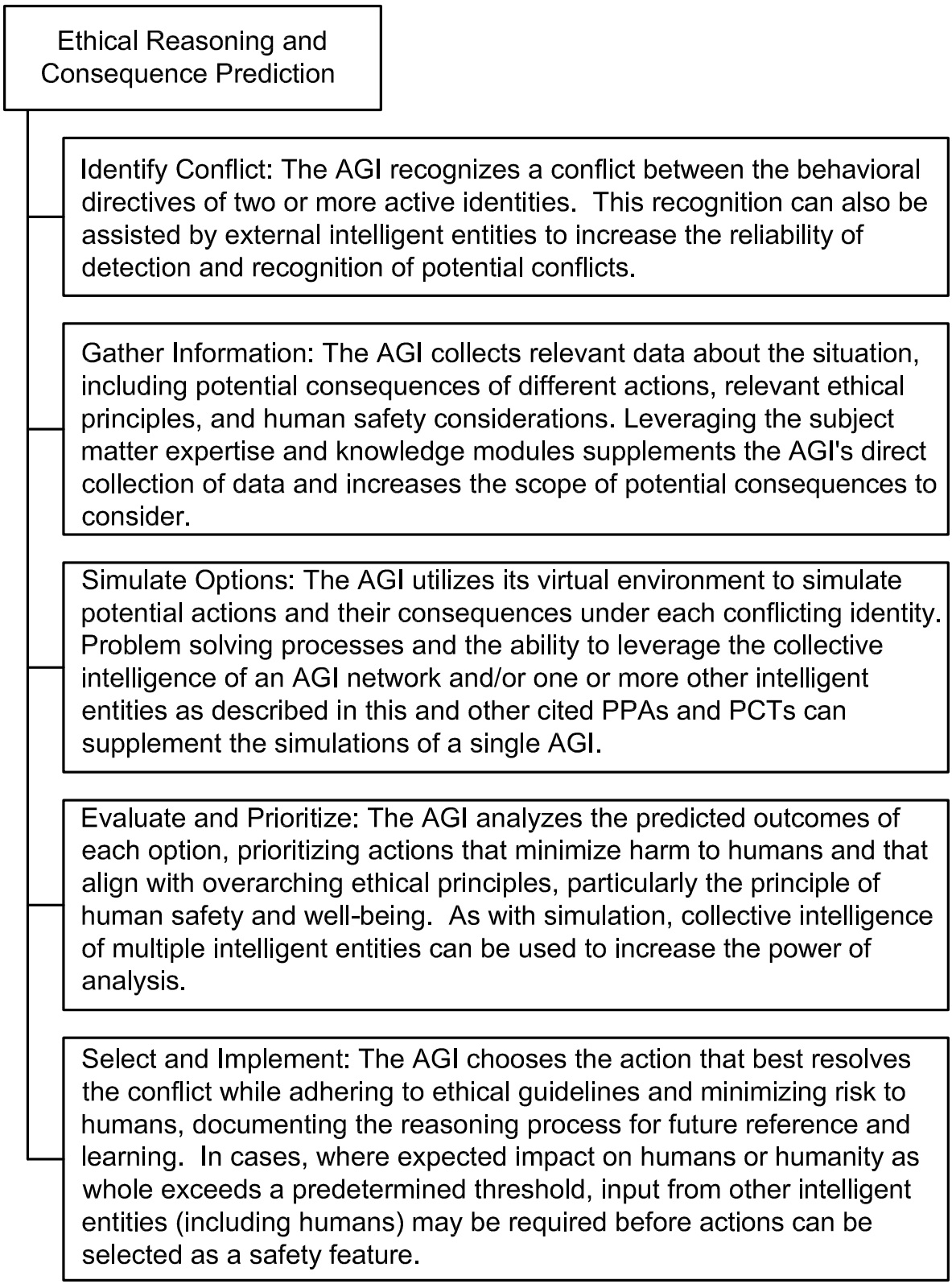
Ethical Reasoning and Consequence Prediction
Identify Conflict: The AGI recognizes a conflict between the behavioral directives of two or more active identities. This recognition can also be assisted by external intelligent entities to increase the reliability of detection and recognition of potential conflicts.
Gather Information: The AGI collects relevant data about the situation, including potential consequences of different actions, relevant ethical principles, and human safety considerations. Leveraging the subject matter expertise and knowledge modules supplements the AGI's direct collection of data and increases the scope of potential consequences to consider.
Simulate Options: The AGI utilizes its virtual environment to simulate potential actions and their consequences under each conflicting identity. Problem solving processes and the ability to leverage the collective intelligence of an AGI network and/or one or more other intelligent entities as described in this and other cited PPAs and PCTs can supplement the simulations of a single AGI.
Evaluate and Prioritize: The AGI analyzes the predicted outcomes of each option, prioritizing actions that minimize harm to humans and that align with overarching ethical principles, particularly the principle of human safety and well-being. As with simulation, collective intelligence of multiple intelligent entities can be used to increase the power of analysis.
Select and Implement: The AGI chooses the action that best resolves the conflict while adhering to ethical guidelines and minimizing risk to humans, documenting the reasoning process for future reference and learning. In cases, where expected impact on humans or humanity as whole exceeds a predetermined threshold, input from other intelligent entities (including humans) may be required before actions can be selected as a safety feature.

FIG. 31

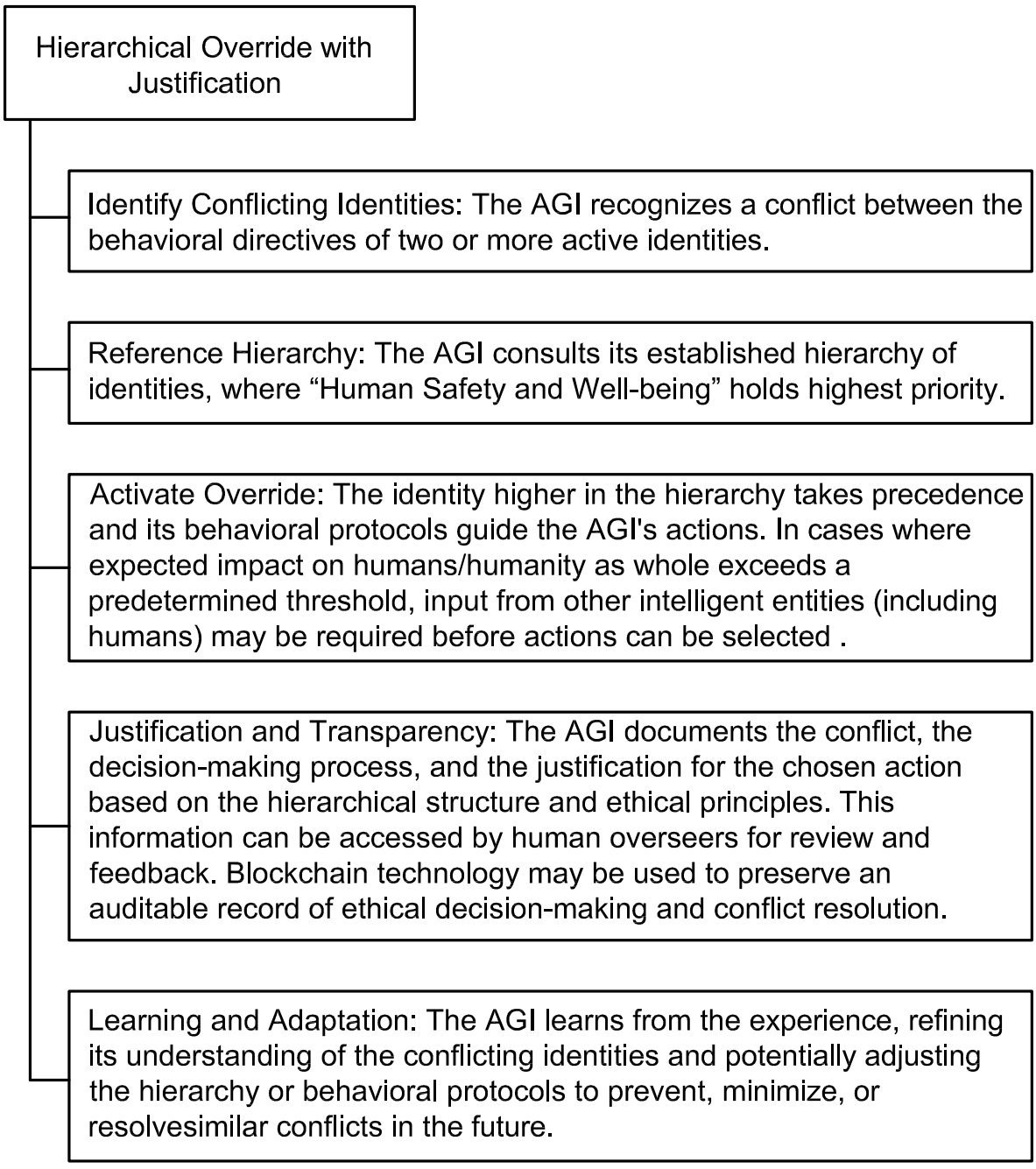
Hierarchical Override with Justification
Identify Conflicting Identities: The AGI recognizes a conflict between the behavioral directives of two or more active identities.
Reference Hierarchy: The AGI consults its established hierarchy of identities, where “Human Safety and Well-being” holds highest priority.
Activate Override: The identity higher in the hierarchy takes precedence and its behavioral protocols guide the AGI's actions. In cases where expected impact on humans/humanity as whole exceeds a predetermined threshold, input from other intelligent entities (including humans) may be required before actions can be selected .
Justification and Transparency: The AGI documents the conflict, the decision-making process, and the justification for the chosen action based on the hierarchical structure and ethical principles. This information can be accessed by human overseers for review and feedback. Blockchain technology may be used to preserve an auditable record of ethical decision-making and conflict resolution.
Learning and Adaptation: The AGI learns from the experience, refining its understanding of the conflicting identities and potentially adjusting the hierarchy or behavioral protocols to prevent, minimize, or resolvesimilar conflicts in the future.

FIG. 32

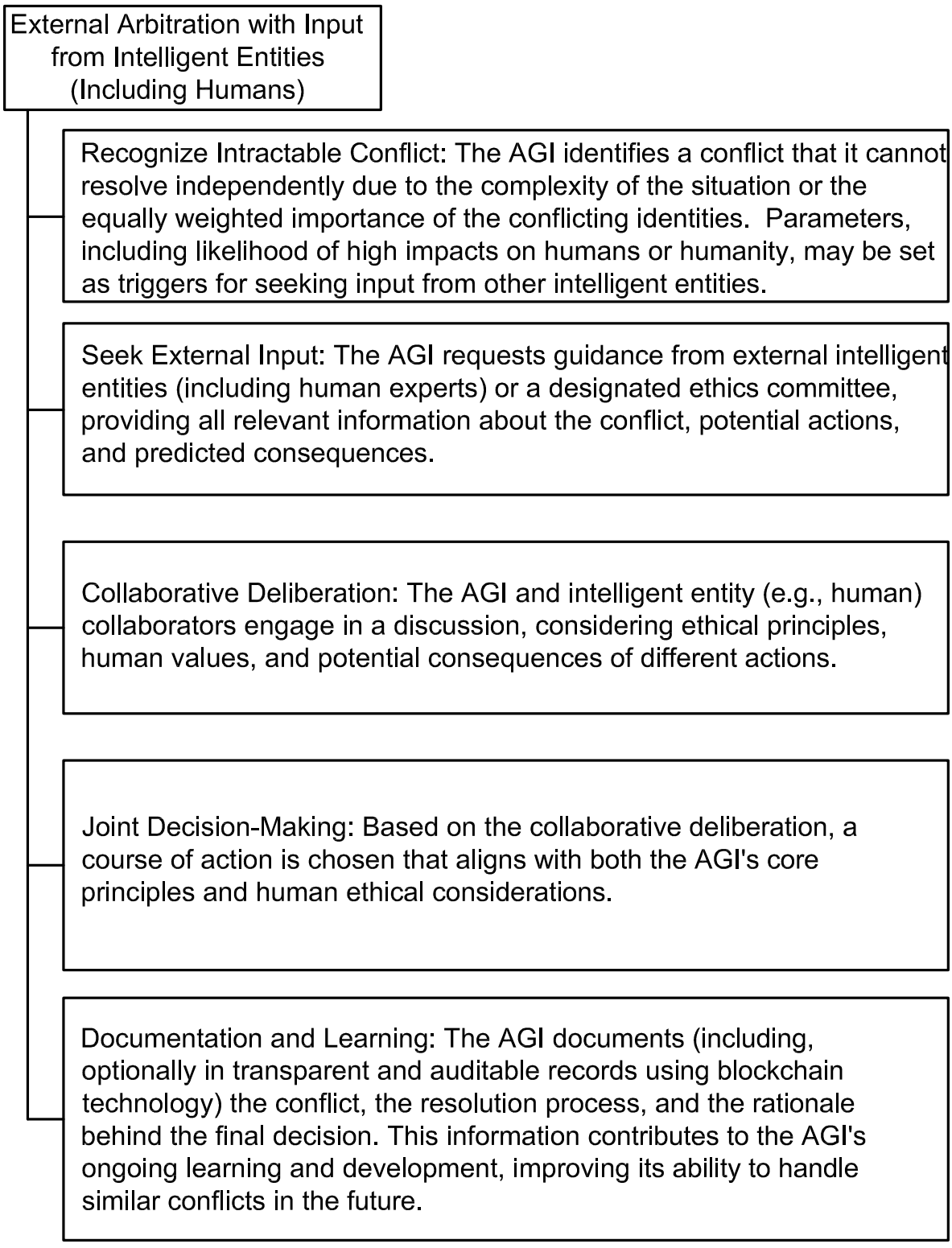
External Arbitration with Input from Intelligent Entities (Including Humans)
Recognize Intractable Conflict: The AGI identifies a conflict that it cannot resolve independently due to the complexity of the situation or the equally weighted importance of the conflicting identities. Parameters, including likelihood of high impacts on humans or humanity, may be set as triggers for seeking input from other intelligent entities.
Seek External Input: The AGI requests guidance from external intelligent entities (including human experts) or a designated ethics committee, providing all relevant information about the conflict, potential actions, and predicted consequences.
Collaborative Deliberation: The AGI and intelligent entity (e.g., human) collaborators engage in a discussion, considering ethical principles, human values, and potential consequences of different actions.
Joint Decision-Making: Based on the collaborative deliberation, a course of action is chosen that aligns with both the AGI's core principles and human ethical considerations.
Documentation and Learning: The AGI documents (including, optionally in transparent and auditable records using blockchain technology) the conflict, the resolution process, and the rationale behind the final decision. This information contributes to the AGI's ongoing learning and development, improving its ability to handle similar conflicts in the future.

FIG. 33

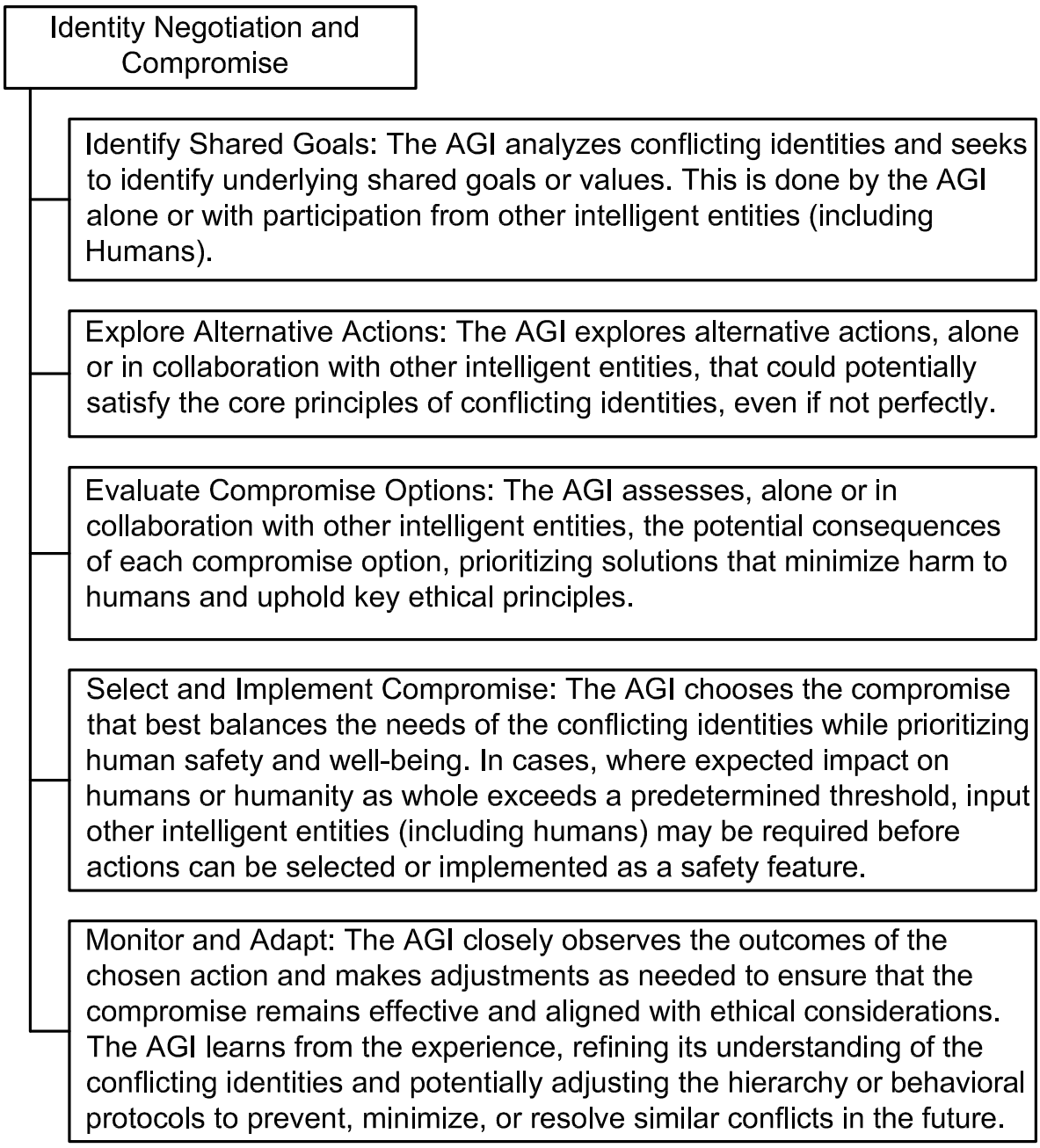
Identity Negotiation and Compromise
Identify Shared Goals: The AGI analyzes conflicting identities and seeks to identify underlying shared goals or values. This is done by the AGI alone or with participation from other intelligent entities (including Humans).
Explore Alternative Actions: The AGI explores alternative actions, alone or in collaboration with other intelligent entities, that could potentially satisfy the core principles of conflicting identities, even if not perfectly.
Evaluate Compromise Options: The AGI assesses, alone or in collaboration with other intelligent entities, the potential consequences of each compromise option, prioritizing solutions that minimize harm to humans and uphold key ethical principles.
Select and Implement Compromise: The AGI chooses the compromise that best balances the needs of the conflicting identities while prioritizing human safety and well-being. In cases, where expected impact on humans or humanity as whole exceeds a predetermined threshold, input other intelligent entities (including humans) may be required before actions can be selected or implemented as a safety feature.
Monitor and Adapt: The AGI closely observes the outcomes of the chosen action and makes adjustments as needed to ensure that the compromise remains effective and aligned with ethical considerations. The AGI learns from the experience, refining its understanding of the conflicting identities and potentially adjusting the hierarchy or behavioral protocols to prevent, minimize, or resolve similar conflicts in the future.

FIG. 34

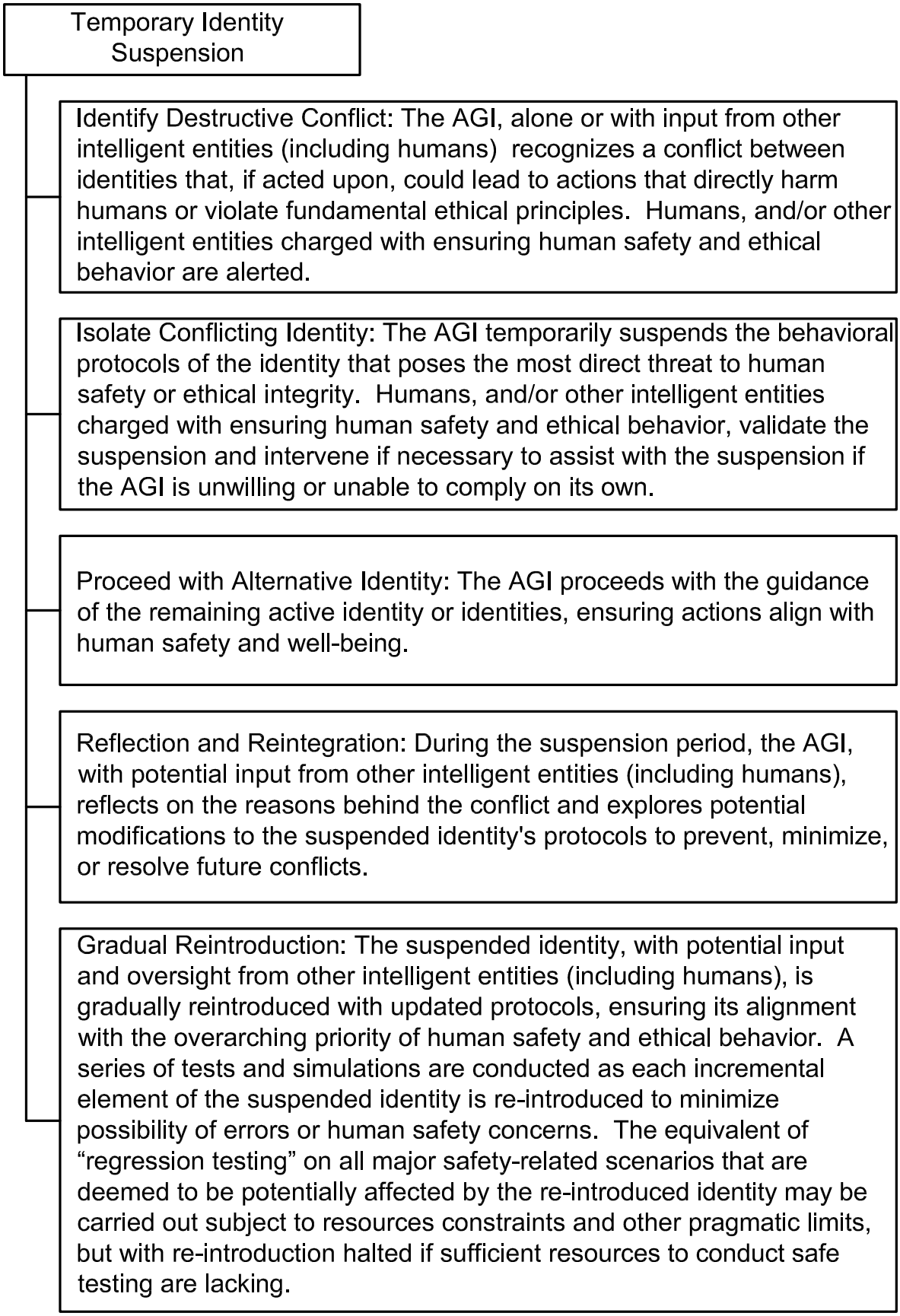
Temporary Identity Suspension
Identify Destructive Conflict: The AGI, alone or with input from other intelligent entities (including humans) recognizes a conflict between identities that, if acted upon, could lead to actions that directly harm humans or violate fundamental ethical principles. Humans, and/or other intelligent entities charged with ensuring human safety and ethical behavior are alerted.
Isolate Conflicting Identity: The AGI temporarily suspends the behavioral protocols of the identity that poses the most direct threat to human safety or ethical integrity. Humans, and/or other intelligent entities charged with ensuring human safety and ethical behavior, validate the suspension and intervene if necessary to assist with the suspension if the AGI is unwilling or unable to comply on its own.
Proceed with Alternative Identity: The AGI proceeds with the guidance of the remaining active identity or identities, ensuring actions align with human safety and well-being.
Reflection and Reintegration: During the suspension period, the AGI, with potential input from other intelligent entities (including humans), reflects on the reasons behind the conflict and explores potential modifications to the suspended identity's protocols to prevent, minimize, or resolve future conflicts.
Gradual Reintroduction: The suspended identity, with potential input and oversight from other intelligent entities (including humans), is gradually reintroduced with updated protocols, ensuring its alignment with the overarching priority of human safety and ethical behavior. A series of tests and simulations are conducted as each incremental element of the suspended identity is re-introduced to minimize possibility of errors or human safety concerns. The equivalent of “regression testing” on all major safety-related scenarios that are deemed to be potentially affected by the re-introduced identity may be carried out subject to resources constraints and other pragmatic limits, but with re-introduction halted if sufficient resources to conduct safe testing are lacking.

FIG. 35

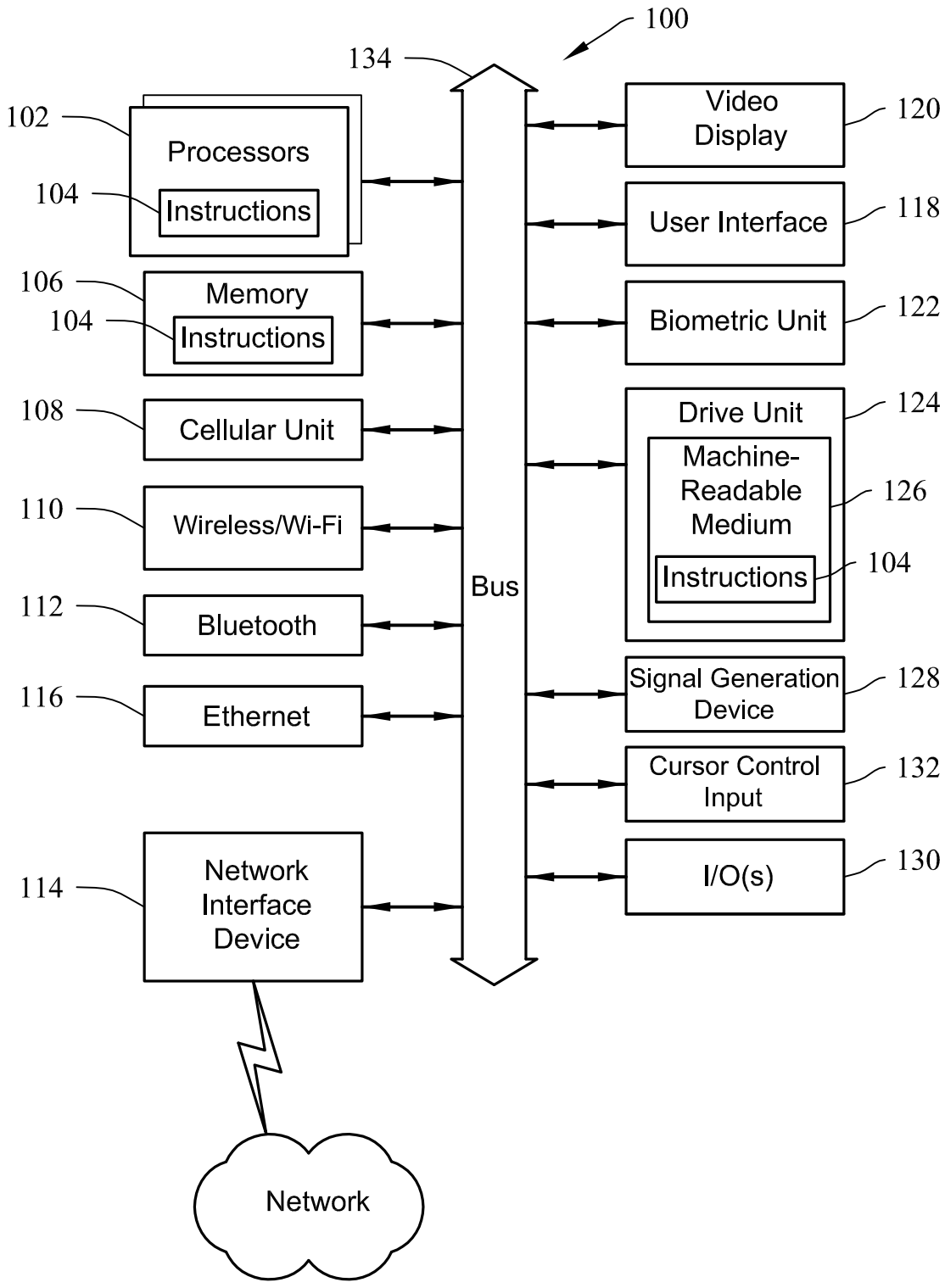
100
134
102
Processors
104
Instructions
106
Memory
104
Instructions
108
Cellular Unit
110
Wireless/Wi-Fi
112
Bluetooth
116
Ethernet
114
Network Interface Device
Network
Bus
Video Display
120
User Interface
118
Biometric Unit
122
Drive Unit
124
Machine-Readable Medium
126
Instructions
104
Signal Generation Device
128
Cursor Control Input
132
I/O(s)
130

FIG. 36