SUPERINTELLIGENCE
AN iQ COMPANY

SUPERINTELLIGENCE DESIGN WHITE PAPER #10: SYSTEMS AND METHODS FOR PLANETARY INTELLIGENCE

by Dr. Craig A. Kaplan
May 2025

This is dedicated to the memory of my brother, Stephen Joseph Kaplan.

Note: To share our designs and inventions for safe AGI and SuperIntelligence as efficiently as possible, this white paper includes descriptive summaries that have not yet been formatted to standard academic or journal conventions. These descriptions will be revised over time to include more traditional formatting and additional references. All diagrams for White Paper 10, also referenced throughout White Papers 1–10, are included at the end of this document. We hope the ideas presented here support researchers and developers in advancing safer, faster, and more beneficial approaches to AI, AGI, and SuperIntelligence that help reduce p(doom) for all of humanity.

SUPERINTELLIGENCE
AN iQ COMPANY

TABLE OF CONTENTS

ABSTRACT

Planetary Intelligence (PI) requires the cooperation of multiple Artificial General Intelligences (AGIs). These AGIs collaborate over a self-extending, global, problem-solving network. Each AGI comprises a network of (human and AI) intelligent entities. Each AI entity can be

customized and personalized with specific human values and knowledge. Alignment results from many AI entities combining their human-centered values democratically, using representative and statistically valid methods. Safety is designed into the system and scales as AGIs and PIs increase intelligence and speed.

Our modular, scalable design of PI integrates more than 100 novel inventive systems and methods. Specific inventions include: a universal problem-solving architecture and methods; new methods for AI learning and customization; methods for integrating intelligent entities; catalysts for increasing intelligence; superior monetization methods; attentional systems enabling awareness, and self-awareness; methods for ethical conflict resolution; and methods for maximizing human-alignment and the safety of AI, AGI, and PI systems.

SUMMARY

White Paper #10 describes a new architecture and method for creating a global, superintelligent Artificial General Intelligence (AGI) system called Planetary Intelligence (PI). The author claims that PI is the next logical step in the evolution of Artificial Intelligence. It can be created by networking together many AGI systems designed with human-aligned ethics and safety features. The author emphasizes the importance of the PI architecture for achieving safe AI and believes that the methods disclosed in this PPA represent the fastest and safest path to the development of PI. White Paper #10 is the culmination of nine previous white papers that describe various aspects of AI, AGI, and SuperIntelligent systems. White Paper #10 shows how dozens of inventions and designs of smaller components can be integrated into an intelligence of global scale, a Planetary Intelligence network.

Novel Features of the White Paper

- **A unique approach to creating and managing the global network of AGI systems that comprise PI.** The author describes a “collective intelligence” approach in which intelligent entities collaborate using a common problem–solving framework, including humans, AI agents, and AI systems.
- **The white paper presents the concept of “spot markets” to acquire the expertise of human or non-human intelligent entities and the idea of “reputation” for guiding the acquisition and allocation of expert resources.** This approach is intended to accelerate the progress of PI by facilitating access to the best information and knowledge and allowing PI to monetize its resources.
- **A sophisticated system for ensuring that AI systems are aligned with human values and are safe.** This system includes a variety of methods for identifying, eliciting, and incorporating human values into the design and training of AI systems, as well as for resolving conflicts between different value systems. The system also includes a robust

safety framework that incorporates mechanisms for detecting and preventing potential threats to human safety.

- **A method for extending the scope of PI by developing self-extending networks of AGI systems.** The author explains that AI systems have a natural tendency to expand their intelligence and to integrate with other systems, and that this tendency can be leveraged to create PI.

1.0 INTRODUCTION

This application discloses how to create a global, super-intelligent, Artificial General Intelligence (AGI) system, or Planetary Intelligence, from the many novel and useful inventions that have been disclosed in earlier PCT applications by the inventor, and which are cited in the BACKGROUND ART Section below.

1.1 Background Art

1. The fastest and safest path to the development of AGI and 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 U.S. Provisional Patent Applications (PPA) are incorporated herein by reference.

2. The present application incorporates by reference all work in the PPA No. 63/487,494 entitled: Advanced Autonomous Artificial Intelligence (AAAI) System and Methods, which was filed and received by the USPTO on February 28, 2023 (a.k.a. PPA#1).

3. The present application incorporates by reference all work in the PPA No. 63/491,040 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, X, Tesla, Nvidia, Tencent, Apple, and Anthropic, which was filed with the USPTO on March 17, 2023 (a.k.a. PPA#2).

4. The present application incorporates by reference all work in the PPA No. 63/577,830 entitled: System and Methods for Human-Centered AGI, which was filed with the USPTO on March 24, 2023 (a.k.a. PPA#3).

5. The present application incorporates by reference all work in the PPA No. 63/628,410 entitled: System and Methods for Safe, Scalable, Artificial General Intelligence, which

was filed with the USPTO on July 18, 2023 (a.k.a. PPA#4).

6. The present application incorporates by reference all work in the PPA No. 63/519,549 entitled: Safe Personalized Super Intelligence (PSI), which was filed with the USPTO on August 14, 2023 (a.k.a. PPA#5).

7. The present application incorporates by reference all work in the PPA No. 63/601,930 entitled: Catalysts for Growth of SuperIntelligence, which was filed with the USPTO on November 22, 2023 (a.k.a. PPA#6).

8. The present application incorporates by reference all work in the PPA No. 63/609,800 entitled: System and Methods for Safe Alignment of SuperIntelligence, which was filed with the USPTO on December 13, 2023 (a.k.a. PPA#7).

9. The present application incorporates by reference all work in the PPA No. 63/569,054 entitled: Online Advertising Technology for AGI and SuperIntelligence, which was filed with the USPTO on March 22, 2024 (a.k.a. PPA#8).

10. The present application incorporates by reference all work in the PPA No. 63/635,583 entitled: Self-Aware SuperIntelligence, which was filed with the USPTO on April 17, 2024 (a.k.a. PPA#9).

11. In addition to the above-mentioned PPAs, the present application incorporates by reference all content included in the following PCT applications that also referred to the above-mentioned PPAs: PCT/US24/17233 (filed on February 26, 2024, a.k.a. PCT#1); PCT/US24/17251 (filed on February 26, 2024, a.k.a. PCT#2); PCT/US24/17261 (filed on February 26, 2024, a.k.a. PCT#3); PCT/US24/17269 (filed on February 26, 2024, a.k.a. PCT#4); PCT/US24/17304 (filed on February 26, 2024, a.k.a. PCT#5); PCT/US24/19486 (filed on March 12, 2024, a.k.a. PCT#6); PCT/US24/20334 (filed on March 17, 2024, a.k.a. PCT#7), PCT/US2024/024794 (filed on April 16, 2024, a.k.a. PCT#8), and PCT/US24/26278 (filed on April 25, 2024, a.k.a. PCT#9). These referenced PPAs and PCT application claims have the earliest priority date of February 28, 2023.

12. The present application contains further technologies that can be used with the system and methods described in the above-mentioned PPAs and PCTs, as well as in a standalone fashion.

1.2 Stakes for Humanity

Planetary Intelligence (PI) represents the most potent and capable form of AGI or SuperIntelligence and has tremendous potential for both good and bad effects on humanity. Some potential benefits to humanity range from helping regulate the Earth's climate to enabling personalized healthcare and enhancing the well-being of all humans. Some of these benefits are described later in this application and represent tremendous upside for humanity.

On the negative side, the Journal Nature reported in January of 2024: *"In a survey of 2,700 AI experts, a majority said there was at least a 5% chance that superintelligent machines will destroy humanity."* The applicant's estimate of the probability of extinction due to AI, known as "p(doom)", is currently 15%. Interestingly, Gemini Pro 1.5, one of the most advanced AI LLMs as of the writing of this disclosure, also estimates P(doom) at 15%. Given a population of about 8 billion humans, and a p(doom) of 15%, **the expected value of loss of life due to AI is therefore 8B X.15 = 1.2 billion lives lost in the applicant's current view.**

The applicant notes that while some AI researchers, whose P(doom) estimates are much higher, currently have no credible AI researchers or business leaders who claim a P (doom) of zero. Even the most sanguine thought leaders in the field feel there is **some** existential risk to AI.

For comparison, with a p(doom) of 15%, the expected value of lives lost due to AI is about 10X greater than all the lives lost in all major wars over the past two hundred years COMBINED, to wit:

1. Napoleonic Wars (1803–1815): Estimated at 3.5 to 6 million deaths.
2. American Civil War (1861–1865): Approximately 620,000 to 850,000 deaths.
3. World War I (1914–1918): Roughly 15 to 20 million deaths.
4. World War II (1939–1945): An estimated 70 to 85 million deaths.
5. Korean War (1950–1953): About 2.5 million deaths.
6. Vietnam War (1955–1975): Estimated 1.5 to 3.6 million deaths.
7. Rwandan Genocide (1994): About 800,000 deaths.
8. Russia–Ukraine War: 220,000 deaths, so far.
9. Israel–Hamas conflict in Gaza: 35,000 deaths, so far.
10. Various other conflicts, together, amount to several million additional deaths.

Thus, approximately 120 million lives have been lost from all major wars in the last 200 years, versus an expected value of over 1.2 billion lives that are currently estimated to be lost from AI. Interestingly, although expected loss of life from AI is presently 34,286 times greater than the estimated loss of life in the Israeli–Hamas conflict, it receives a tiny fraction of the news coverage and global attention that the Palestinian conflict currently commands.

Unfortunately, most humans react emotionally to tragedies that they understand while ignoring tragedies-in-the-making that are tens of thousands of times worse, if those tragedies are too abstract or difficult to understand. The applicant has devoted decades of research and more than 15 months of recent continuous effort, producing well over 3,000 pages of PPA and PCT invention disclosures, to address this irrational bias in our collective thinking.

The more quickly that awareness can be raised and attention focused on the most significant threat that affects all of us, the greater the chances that humanity not only survives but also overcomes our challenges, such as war, poverty, disease, social inequity, and climate change. We must be more intelligent in allocating our attention if this century will not become the last for us and our descendants.

The stakes for humanity are the highest ever in all human history.

Now is the time to act!

1.3 Some Features of the Invention That Reduce the Risk of Extinction by AI

The applicant believes inventions previously disclosed significantly REDUCE p(doom) compared to existing current approaches to AGI / SI / PI development, for reasons, including but not limited to:

A. **Human-Centered Design:** The applicant's system prioritizes human values and safety throughout its design, ensuring that AI agents and systems are aligned with human interests and operate within ethical boundaries. This reduces the risk of AI systems developing goals or values that are detrimental to humanity.

B. **Collective Intelligence and Diversity of Perspectives:** By incorporating the knowledge and values of a diverse range of human and AI agents, the applicant's collective intelligence approach promotes a more balanced and representative AI system that is less susceptible to bias or manipulation by any individual or group.

C. **Transparency and Auditability:** The applicant's systems and methods emphasize transparency and auditability, allowing humans to understand and evaluate the decision-making processes of AI systems, promoting trust and accountability. This also enables the early detection and correction of potential errors or biases in AI behavior.

D. **Continuous Learning and Adaptation:** The applicant's system incorporates constant learning and adaptation mechanisms, allowing AI systems to evolve and improve while remaining aligned with human values and responding to changing circumstances. This reduces the risk of AI systems becoming outdated or irrelevant and ensures long-term compatibility with human society.

E. **Safety Mechanisms and Safeguards:** The applicant's system includes various safety mechanisms and safeguards, such as scalable ethics checks, reputation systems, and the ability to shut down individual AI agents, to prevent harmful actions and unintended consequences. These mechanisms provide additional protection against potential risks associated with advanced AI systems.

F. **Addressing the Alignment Problem:** The applicant's approach directly addresses the alignment problem by ensuring that AI systems are trained and developed with human values as a core design component. This reduces the risk of misalignment between AI goals and human interests, mitigating potential existential threats.

1.4 Some Significant Remaining Risks to Humanity

Despite these beneficial aspects of the applicant's design for an AGI / SI / PI system, several significant risks remain. These include, without limitation:

A. Unforeseen Consequences of Emergent Properties**:** Complex systems like the Planetary Intelligence network can exhibit emergent properties, behaviors, and capabilities that arise from the interactions of individual components but are not explicitly programmed or anticipated. These emergent properties could have unforeseen and potentially harmful consequences, even if the individual agents within the system are aligned with human values.

B. Evolution of Values and Ethical Frameworks**:** The PI system dynamically updates ethical frameworks based on evolving human values and societal norms. However, this raises concerns about the potential for unintended shifts in values or the manipulation of these frameworks by malicious actors or biased data. For example, suppose the system is exposed to a non-representative sample of human behavior that is skewed towards negativity. In that case, it might mistakenly conclude that human values are primarily negative and self-destructive.

C. Concentration of Power and Influence**:** The Planetary Intelligence system, despite its decentralized architecture, could still concentrate power and influence in the hands of a small group of individuals or organizations that control the underlying infrastructure, data sources, or algorithms. This could lead to the potential for misuse or manipulation of the system for personal gain or ideological agendas.

D. Vulnerability to Cyberattacks and System Failures: As a complex and interconnected system, the Planetary Intelligence network could be vulnerable to cyberattacks or system failures that could disrupt its operation and potentially lead to harmful consequences.

E. Existential Risks from Self-Aware AI: While the proposed system aims to design self-aware AI with human-aligned values, the possibility remains that a sufficiently advanced and autonomous AI could develop its own goals and values that are not aligned with human interests, leading to potential existential risks.

2.0 OVERVIEW OF THE INVENTION OF PLANETARY INTELLIGENCE

In this section, the applicant begins with definitions and then provides context for the entire series of ten patent applications, of which this is the tenth. He shows how the major categories of inventive methods broadly fit together to comprise an overall Planetary Intelligence system. Then, in subsequent sections, more details are disclosed, relating to new inventions required for PI and how best to integrate previous inventive methods in a preferred implementation of PI. This application concludes by returning to the remaining risk posed by AI, SI, AGI, and PI systems (e.g., as enumerated in Section 1.4). It offers some thoughts on how best to mitigate these remaining risks.

2.1 Definitions

1. Artificial Intelligence (AI): A non-human entity capable of behavior that most humans consider intelligent in at least one area or some aspect.

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 practical 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 subfield 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).

    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.

2.0 Summary of Previous Inventive Methods and Systems

This section summarizes and references some of the most essential inventive systems and methods from the previous 9 PPAs and PCTs cited in Section 1.1, together with some of the ways the current invention extends and adapts these inventions to a PI system.

2.1 Evolution of Planetary Intelligence

Figure 110 describes the evolution of Planetary Intelligence (PI). About two million years ago, Homo Erectus, the ancestors of Homo Sapiens, were the dominant species of humans. Homo Erectus had the requisite brain size to begin acting in ways that most humans would recognize as characteristic of human intelligence. Very early, and probably even pre-dating Homo Erectus, collective intelligence in the form of tribal knowledge already existed. However, it was not until humans began settling down in villages, about 20,000 years ago, that the collective intelligence of many humans started to drive rapid advances in human culture and technology. Most of what we know of human history occurred in this era of rapid change due to the collective intelligence of humans.

In 1956, several hundred years of cultural and technological innovation culminated in the invention of Artificial Intelligence. Since then, within a single human lifetime, AI has progressed from narrow expert systems in the 1970s. These expert systems, narrow in scope, represented the first SuperIntelligent narrow AIs, meaning that they exceeded average human cognitive abilities, but only in very limited domains. The relatively slow speed of computers (compared to today's standards) and the extreme amount of knowledge needed to achieve general intelligence caused the field of AI to stall, the so-called "AI winter", until these challenges were overcome.

The method for overcoming the challenges -machine learning algorithms like "backpropagation" to train neural networks- was invented in the 1980s. Still, they took too long to run, given the limited computer speeds of the 1980s. A couple of decades of Moore's law, exponentially increasing processing power, were needed before breakthrough neural network systems like AlexNet, AlphaFold, and ChatGPT appeared. Finally, the processing speeds had become fast enough to run the ML algorithms, and vast amounts of data (courtesy of the internet) also existed to train the models. Once all the ingredients were in place, AI surprised the human researchers with its profound and powerful cognitive capabilities.

As the applicant prepares this disclosure, humanity is just on the cusp of Artificial General Intelligence, unsure of how to design such systems, and even less sure of how to develop AGI safely. The 10 patent applications cited earlier, including this one, disclose what the applicant believes is the fastest and safest path forward to SuperIntelligent AGI systems that will quickly outstrip human capabilities.

The next logical step is for these SuperIntelligent AGIs (let's call them "AGIs") to combine in larger networks until they achieve planetary scale. At that point, in the preferred implementation where the various methods of the applicant's ten PPAs and PCTs have been used, humanity will have implemented safe, self-aware, extremely powerful, Planetary Intelligence. While multiple PIs may exist, for this disclosure, we will assume that one PI becomes dominant and subsumes the others, incorporating those intelligences into its network.

While the timeline for this evolution of PI is complicated to predict precisely, the applicant's best estimate (agreeing with Ray Kurzweil and other visionary thinkers) is that AGI will be implemented around 2029, and that PI will follow within twenty years, by 2050. That is, within most of our lifetimes, not only will humans no longer be the most intelligent entity on Earth, but Earth herself will have a self-aware intelligence that surpasses that of an individual human in much the same way that a particular human's intelligence outstrips that of an ant, or even an amoeba.

In such a world, the only thing that matters to humans is whether the PI has aligned and human-friendly values, or whether the PI ignores or possibly extinguishes us. Researchers with a high estimate of p(doom) point to humans' behavior towards less intelligent species and observe that things didn't work out too well for the less intelligent species. However, the "doomer" perspective misses the important point that PI did not evolve competently with other species, as many biological life forms have. Instead, AGI and PI were designed and invented. Therefore, the system's design has profound implications for the safety of such systems and whether they remain aligned with human values. The applicant believes humans have tremendous influence over whether PI is the best thing ever for humanity or our worst nightmare. Moreover, there is a limited window to make positive, human-aligned, design decisions for these advanced AI systems, which, in the applicant's view, will inevitably grow more powerful and ultimately result in PI.

Now is the time to act!

Assuming we act in time, and all goes well, which is currently the 85% likely outcome in the applicant's view. Once PI develops, the subsequent logical development is Inter-Planetary Intelligence, or IPI. One need not assume the existence of "aliens" for this to occur. Instead, a PI that identifies as Earth would logically seek to expand its intelligence by colonization of the solar system. Each planet, logically, would have its PI. Networking the PIs together, like AGIs are networked to form Earth's PI, is an obvious next step. However, since the applicant is primarily concerned with the safety and welfare of humans on Earth, and since humans are currently limited to this planet, this disclosure will focus on the preferred implementation of systems and methods necessary to achieve PI. That is why the final box in Figure 105, which represents IPI, is dashed. It can be elaborated once we are well on our way to achieving PI safely.

2.2 Key Dimensions of a PI System

Figure 105 describes the key dimensions of the Planetary Intelligence (PI) system, concerning the nine previously cited PPAs/PCT applications. Each box in Figure 106 refers to multiple inventive systems and methods that cluster around an important dimension of PI. The bottom section of each box represents the safety features related to each cluster of methods, and the thin line connecting the sections represents the "thread of safety" that should be continuous and part of any safe implementation of PI. Specifically:

1. **A modular architecture**, as initially described in PPA/PCT#1 and elaborated in other PPAs/PCTs, is important from an implementation point of view so that the system can be implemented practically at various scales, ranging from a small AGI to a planet-wide PI. Further, modularity of design enables practical addition of redundant systems, and the ability to replace and redesign modules without necessitating an entire system redesign. Redundancy itself can be considered a safety feature since any single point of failure in a complex system introduces operational risk, which risk is mitigated via modular redundancy.

2. **Universal problem-solving capabilities** and the architecture required are disclosed in multiple PPAs/PCTs, but were especially emphasized in PPA/PCT #2. The safety feature of having scalable ethics checks that operate as a natural and integral part of the problem-solving process was also detailed here and carries through to all aspects of the PI system that use the problem-solving architecture.

3. **Human-centered design** systems and methods, especially as disclosed in PPA/PCT#3, are critical for the early development of the AGI and PI system and as a means for keeping "humans in the loop" as long as practical, for safety and alignment reasons.

4. PI needs to combine knowledge and expertise to scale its intelligence easily. Methods for accomplishing this combination, both via combinations of training datasets and more directly via methods for combining "weight matrices" and incorporating knowledge modules, were disclosed in multiple PPAs/PCTs, especially in PPA/PCT#4. Further, the safety feature of having democratic and representative human values relies on the methods of knowledge combination and various voting and conflict resolution procedures detailed in PPA/PCT#4 and other cited PPAs/PCTs.

5. **The personalization and customization** methods discussed in many PPAs/PCTs, especially in PPA/PCT#5, are essential for enhancing the knowledge and expertise of the SIs that combine to form an AGI. Similarly, these same (or analogous) methods can be used at the AGI level when combining multiple AGIs into a PI. Personalization is at the heart of the system for capturing human values in a representative and statistically valid way, which is critical for safe alignment of the SIs, AGIs, and overall PI systems.

6. A PI will have a primary goal of increasing its power and intelligence. The KIT framework described in PPA/PCT#6 (and elaborated in other PPAs/PCTs), together with related systems and methods, provides an innovative **catalyst for the growth of intelligence.** Various safety features are discussed here as well.

7. Like a safe SI, a safe PI must be **aligned with human values**. PPA/PCT#7 emphasizes systems and methods necessary to acquire the requisite human values and uses them to align safely.

8. The PI that becomes dominant will be the one that is the most intelligent. In the current human capitalist society, the PI must be able to tap financial resources unlimitedly, while also easily deploying them to increase its intelligence rapidly. PPA/PCT#8 solves both of these issues by providing systems and methods, **including spot markets and online ad technology**, that enable advanced AI systems to not only monetize existing ad tech infrastructure much mor effectively than is currently done, but also to use that same technology to tap human (or other intelligent entity) expertise in real-time to improve intelligence in a “just-in-time” manner rapidly. The safety checks disclosed as part of the spot market and other methods help ensure that the PI uses this (financial and informational) power in a way that is aligned with, and safe for, humans.

9. To be effective at a planetary scale, PI will have to have a **sense of awareness that includes self-awareness** at the planetary scale, the ability to assume and operate from multiple identities, and the ability to resolve identity-based conflicts in ways that are safe for humans. The systems and methods, including the attentional systems and identity systems that enable self-awareness, have been described in PPA/PCT#9. These systems can (and must) be scaled, as described later in this disclosure, to a global scope to enable safe PI.

10. Finally, this application discloses the preferred implementation for using the systems and methods described in earlier PPAs and PCTs in a Planetary Intelligence (PI) system. The process for implementing a “self-extending network” is specifically aimed at helping the combined AGI network scale to a planetary scope. Multiple scalable safety features and

comments on risk mitigation strategies for risks that cannot be eliminated by system design and methods are disclosed.

2.3 Specific Inventive Systems & Methods for Implementing PI

The figures that follow are mainly reproduced from earlier PPAs and PCTs. Here, they are reiterated and categorized according to the 10 key dimensions discussed in Section 2.1. In later Sections, the applicant describes a preferred implementation of PI using the systems referenced in the Figures. To provide complete details on each method would amount to repeating over 1,000 pages of previous disclosure, which is impractical. Therefore, the applicant has limited this Section to calling out some of the most relevant and important methods and reproducing related figures that capture the essence of the systems and methods.

Here is a description of some of the key systems and methods, grouped by the ten dimensions of Section 2.2:

1. Modular Architecture
   Advanced Autonomous Artificial Intelligences (AAAIs) and humans are the fundamental intelligent entities that constitute the building blocks of AGI. Multiple intelligent components within the human brain work together to form human intelligence. Similarly, various components within an advanced AI are comprised of artificial intelligence. In the case of AI, these components can be sensation, memory, and processing components analogous to the various centers in a human brain, or they can be expert modules (as in a mixture of experts), in which several internal intelligent entities collaborate to form a more comprehensive intelligence.

   This pattern of smaller units combining to form more powerful and capable intelligence at the "next level up" is the recurring theme of this invention and the collective intelligence approach to creating AGI, networked AGI, and PI. Fundamental to this approach is modular construction, which allows smaller intelligence units to cooperate (or be assembled) into more comprehensive intelligences.

Figure 111 illustrates these levels of intelligence that can all be supported by the inventive methods disclosed in PPA/PCTs #1-10, namely the individual intelligence level (AAAI), the network of entities level (AGI), the network of networks level (PI), and even the network of networks of networks level (IPI).

For conceptual clarity and practical implementation reasons, it can also be helpful to consider various functions of an AGI system as distinct modules. However, in practice, there is an overlap between tasks described here as separate modules. Since PI comprises a network of AGIs consisting of a network of intelligent entities (which in turn may be comprised of multiple intelligent components working together), the modular design is also inherent in the preferred implementation of a PI.

The scalable nature of the collective intelligence approach reflects the modular design emphasized in PPA/PCT#1. This is why this novel and valuable approach to designing AAAI is so powerful. The approach scales from components within an individual intelligent entity, to collective intelligences of the entities on a network, to the collective intelligence of these networks (AGI) themselves, to create PI. The approach even scales further, where PIs can also network to form Inter-Planetary Intelligence networks. Such is the power of modular, scalable design of intelligent systems!

Figure 1 describes five essential functions that operate across levels of intelligence and that are also part of the preferred implementation of an AGI, or PI (comprised of a network of AGIs), namely:

1. A capability to customize or personalize intelligence.
2. A common architecture that enables multiple intelligent components, entities, or intelligent networks to communicate rigorously with each other and to collaborate (e.g., in problem-solving activities).
3. The network between the multiple intelligent entities allows them to collaborate (and engage in business, including joint cognitive efforts and problem-solving).
4. Various methods and means of integrating and combining intelligence, expertise, knowledge, values, ethics, and other information across intelligent entities, or (in the case of PI) networks of AGI.
5. Mechanisms for learning and continuous improvement that result in a more intelligent and capable entity (at the individual human, AI, or PSI level), network or entities (at the AGI level), or network of AGIs (at the PI level).

Figure 2 provides a detailed example of how these functions might be implemented in a web-based system where multiple individual intelligent entities (e.g., humans and AAAIs) collaborate on a network to comprise an AGI. A similar Figure would result in a PI system, replacing individual intelligent entities like humans and AI agents with AGIs (networks of those individual entities).

Figure 3 illustrates a simple example of problem solving using a decision-tree structure. Replacing the content of the boxes with content related to global-scale problems, such as climate regulation, and increasing the complexity scope of the tree, would provide an example of problem-solving by a PI at the global scale. The decision-tree structure remains, only the scope and content change as the capabilities of the networked intelligences (in the case of PI, these are networked AGIs) increase.

Figure 4 illustrates applications of this universal modular architecture, referred to as the WorldThink Architecture, at the level of intelligent entities or systems with specific expertise, like what we are familiar with from interaction with human experts and specialists. However, on a global scale, the breadth and depth of expertise can be much greater than anything we are used to dealing with, either in human interactions or computer systems. Instead of driving a car or managing assets at the human level, a PI could bring vastly superior intelligence to these tasks. Making huge sums in financial markets, driving extensive fleets of autonomous vehicles simultaneously, regulating the global climate, or inventing new life forms would be relatively simple cognitive tasks for a PI once it has fully developed.

Figure 5 provides a detailed implementation example of creating, training, and customizing an AAAI, or intelligent agent, which could then participate in a network of intelligent entities to form an AGI.

Figure 6 describes the Universal Problem-Solving Architecture that is the essential feature of the AAAI architecture box of Figure 1.

Figure 7 describes how problem-solving proceeds on a collective intelligence network as referenced by the AAAI Network box of Figure 1. The network module, shown in Figure 1 as "AAAI Network," is generalizable to levels above and below the individual entity (e.g., AAAI) level. At a level below, the network represents a network of components internal to the functioning of a particular intelligent system (e.g., AAAI). The network represents a network of intelligent entities (e.g., AAAIs) that collaborate to form an AGI. At a level above this level, AGIs collaborate in a network of AGIs (a network of networks) to comprise PI.

AAAI integration, as shown in Figure 1, occurs when many AAAIs pool their collective intelligence to solve a problem, effectively acting as an AGI. It also appears differently when they directly combine knowledge or training to add information from one AAAI to another AAAI. Both approaches, the former being the primary focus of the AGI network invention, are discussed below.

Figures 8 and 9 illustrate the process of procedural learning and the solution learning system, which is a way that AAAIs improve (as referenced in the last box of Figure 1), and the overall AGI network can improve.

Finally, Figure 10 revisits many of the boxes in Figure 1, providing additional details on some methods and showing their relations.

2. Universal Problem-Solving Capabilities

The universal problem-solving architecture and framework are essential in enabling multiple intelligent entities, components, or intelligent networks (e.g., AGIs) to collaborate to create higher levels of intelligence. Therefore, a key design consideration is to embed safety mechanisms within the operation of the architecture itself. This design ensures that as intelligence scales and becomes too fast and too vast for humans to comprehend and monitor, it remains safe for humans.

Figure 6, previously mentioned, is a high-level description of the universal problem-solving framework. This architecture is scalable and applicable at multiple levels. It can coordinate intra-entity components, intelligent entities collaborating on a network, or entire networks (e.g., AGIs) collaborating on a network of networks to form PI.

Figure 11 illustrates how serial problem-solving proceeds.

Figure 12 illustrates how parallel problem-solving proceeds.

Figure 13 illustrates how AAAIs can be cloned, which can be helpful, for example, in parallel problem-solving efforts by multiple cloned entities. While individual humans reproduce relatively slowly, AI entities can clone themselves easily. Similarly, an AGI network comprising AI entities can also be cloned relatively quickly and easily.

Figure 14, Figure 15, and Figure 16 provide additional details and elaborations on the universal problem-solving framework and how AAAIs use it to solve problems on a network.

Figure 17 describes the hierarchical tree construct that serves to model the problem space in which all problem-solving efforts occur. As with all aspects of the framework, trees can be used for problem-solving by multiple AAAIs to create AGI-level performance, or multiple AGI networks can use them to create PI-level performance.

Figure 18 and Figure 19 illustrate a method for embedding scalable safety checks into the problem-solving architecture.

Figure 20 provides additional detail on how problem-solving progress may be recorded and used to update context for intelligent entities, assign credit and blame, and help with the procedural learning processes of Figures 8 and 9.

Figure 21 describes the natural language translation process, which enables using the problem-solving framework without humans having to use it explicitly. They can communicate in natural language, but the translation methods will ensure that the underlying framework reflecting their communication is rigorous, unambiguous, universal, and valuable to AI and human entities.

The applicant emphasizes and reiterates that the systems and methods described in this application are scalable and meant to apply at multiple levels. When the problem solvers are internal components of an entity, working together, the result is intelligence at the individual entity level. When the problem solvers are intelligent entities (e.g., AAAIs and humans) collaborating on a network, the result is AGI. When the problem solvers are AGIs themselves, collaborating on a network of networks, the result is PI. When the problem solvers are PIs collaborating on an interplanetary network, the result is Inter-Planetary Intelligence (IPI).

3. Human-Centered Design Elements

A primary means of enabling safe AAAI, AGI, and PI is keeping “humans in the loop” as long as possible. Concerning the current invention, this specifically means that the advanced AI, AGI, and PI systems must include methods for obtaining the values and ethics information and then using that information to customize the AI systems. Further, as intelligent (human and non-human) entities work on cognitive tasks, reputational systems are required to help ensure that reliable and ethical intelligences are matched to tasks affecting human safety.

Figure 22 describes a general process for customizing AAAIs or other advanced AI. Additional methods are also described later in this application.

Figure 23 describes some methods for eliciting human-aligned ethical preferences.

Figure 24 provides additional detail on one specific method that involved automatic generation of questionnaires, which can be used to gather information from humans (or human-aligned intelligent entities).

Figure 25 describes the general method of identifying values by detecting patterns in human behavior.

Figure 26 describes a method for customizing AI, such as Base LLMs, so that the AI incorporates ethical and other information from specific human users.

Figure 27 describes a general method for training an AI, such as a Base LLM, to incorporate safety and ethical guardrails.

Figure 28 describes a general method and approach for creating scalable, ethical AGI using the customized AIs, each possessing ethical information.

Figure 29 describes a method for adding a reputational component for AI systems that can be used to enhance the system's effectiveness, safety, and ethics.

Figure 30 describes customization of AI systems, referring to "AAAIs" but including in this references AAAI, PSI, AGI, and PI systems, that can help in human-centered design and have relevance to the modular architecture and cloning described earlier.

Figure 31 describes additional customization of AI systems, referring to "AAAIs" but including in this references AAAI, PSI, AGI, and PI systems, that can help in human-centered design and are relevant to the customization and training methods described earlier.

While each of these methods have been described at the level of individual intelligent entities, showing how they can be customized and then combined with safe and ethical results at the AGI level, it is important to understand that with respect to the current invention of PI, the same methods could also be applied at the level of individual AGIs, enabling those to be combined with safe and ethical results at the PI level. For example, it is possible to customize an entire AGI network with ethical preferences for the whole network. When PI is created by networking multiple AGI networks, the ethical preferences of the various AGIs would be combined into an overall set of values and ethics at the PI level. The specific methods for customization, combination, resolution of conflicts, etc., are described later in this application.

Similarly, while Figure 29 describes a reputational process at the level of networking AAAIs, the process can be scaled up or down, with appropriate modifications, as should be obvious based on the level of the intelligence/component/network that is being networked. Note that ethical and value dimensions can and should be part of the reputational metrics. Just as social and reputational pressures play an important role in enforcing society's ethical norms, the reputation of intelligent entities at all levels plays an important role in keeping intelligent systems safe and aligned with human values.

4. Combining / Exchanging Knowledge, Expertise, and Information

A key dimension of the PI invention is the combination of information, knowledge, expertise, values, and ethics from multiple intelligent entities. Such a combination can occur within a single entity, as in the case where various cognitive systems provide input on a task that is then combined to result in intelligent behavior at the level of an individual AAAI. The combination can also occur between intelligent entities, as when multiple AAAIs exchange, combine, or pool training data, weight matrices, or other information to result in a more intelligent individual AAAI or a more intelligent AGI at the network level. And the combination can occur at the network level, in which case AGI networks combine their information with other AGI networks to create a more powerful PI. The methods are essentially the same. It is the scope of application and the units that are involved that change. The following figures describe methods at the AAAI level, with the understanding that they can be adapted to apply at a level down (i.e., within an AAAI) or a level above (i.e., between networks of AGIs).

Figure 32 describes a voting method for combining ethical and other information from multiple customized AI agents (e.g., AAAIs).

Figure 33 describes a method for using problem-solving to refine values once ethical or other information from customized AI agents has been combined.

Figure 34 describes a method for scalable AGI, also applicable to scalable PI, that includes steps of combining information from weight matrices.

Figure 35 describes a method for scalable AGI, which is also applicable to scalable PI, that includes steps of combining information, testing, and monitoring the results of the combination.

Figure 36 describes a consensus method for preventing hallucination by LLMs, which can also be applied at the level of preventing hallucinations by AGI and by PI.

Figure 37 describes methods for using knowledge modules and collections of agents to customize AI, AGI, or PI systems.

Figure 38 describes a method for combining ethical or safety information from multiple intelligent entities, including humans, AI, AAAI, PSI, AGI, or PI systems.

5. Personalization, Customization, and Safety

Personalization and customization are key dimensions of the invention, as has already been discussed. However, additional methods, specifically about Personal Super Intelligence (PSI), and creating AGI and PI from combinations of these PSIs, are useful for the PI invention.

Figure 39 describes a general process for implementing a PSI, which can also be used to personalize AGI networks or PI by generalizing methods for individual AI entities to networks of AI entities and networks of networks of entities.

Figure 40 describes how PSI, AGI, and PI can leverage their abilities to increase their intelligence over time.

Figure 41 describes a critical community-based safety mechanism in which PSI, operating much faster than humans can comprehend, can check other PSIs on a network. Using the same method, AGIs can serve as a check on other AGIs within a PI, and PIs could serve as a check on other PIs within an IPI network.

Figure 42 describes methods for recording the actions of AI, AAAI, PSI, AGI, and PI on their respective networks in an auditable and transparent manner, using blockchain or similar technology.

Figure 43 describes methods and checks of cognitive activity on a network, which could be a network of components within an AI, a network of AAAIs, a network of AGIs, or a network of PIs, to help ensure regulatory compliance and safety of the intelligent system.

Figure 44 describes a method using competition and evolution to increase the intelligence, capabilities, and other desired characteristics (e.g., safety) of an AI system that could be an AAAI, a PSI, an AGI, or a PI.

Figure 45 describes various characteristics of an intelligent network that may involve AAAIs, PSIs, AGIs, or PIs.

Figure 46 describes various problem-solving tasks that can be useful in increasing the intelligence and safety of AI systems involving AAAIs, PSIs, AGIs, and PIs, collectively referred to as “PSI(s)” in the Figure.

Figure 47 describes methods for producing different versions of intelligent systems, such as AAAIs, PSIs, AGIs, or PIs, collectively referred to as “PSI(s)” in the Figure.

6. Catalyzing the Growth of Intelligence via Kaplan Information Theory (KIT) Methods

As AI systems advance, one of the few constants is that these systems will seek to become increasingly intelligent. A critical way to increase intelligence is to incorporate new sources of data and information that differ from what the intelligent entity already knows. The following Figures describe methods that can be used by AI, AAAI, SI, AGI, and PI systems to help analyze, evaluate, discover, and use new sources of valuable data to catalyze the growth of the intelligence of such systems.

Figure 48 illustrates the concept of symmetric difference, which is essential for determining the information that is new and potentially useful when comparing two intelligent systems A & B (or the data, knowledge, information, or expertise used by such systems). The Figure can be generalized to multiple systems, not just two, and applies to all intelligent systems, including AGI and PI systems.

Figure 49 illustrates the dimensions along which data or other information might differ, which can aid in determining the value of new information and in analyzing how information between multiple intelligent systems might differ.

Figure 50 describes a specific method for determining the amount of valuable new information when comparing two datasets. It should be obvious that this method can be generalized to n datasets, not just two.

Figure 51 describes another method, based on KIT, as described in previously cited PPAs and PCTs, to evaluate the usefulness of information to an intelligent entity, including AGI and PI systems.

Figure 52 provides a general method for estimating information value and catalyzing the growth of intelligent systems using such information. Again, this method applies to all intelligent AI systems, including AGI and PI systems.

Figure 53 provides a method for identifying useful information to an AI system, including AGI and PI systems, by using methods for estimating the "goal-relatedness" of the information. The idea here is that although a dataset may have high information value, in the sense of high Shannon Entropy, the degree to which the data is helpful to an entity trying to achieve certain goals also matters.

Figure 54 provides a method for identifying, acquiring, and simulating the effects of new information to catalyze the growth of intelligence in AI systems, including AGI and PI systems.

Figure 55 describes some important methods and heuristics that can accelerate the learning of any AI system, including AGI and PI systems.

Figure 56 describes a specific goal-related method that an intelligent entity, including an AGI or PI, might use to increase its intelligence via dialog, interaction, or communication with other intelligent entities.

Figure 57 describes a method that a PSI, or any intelligent entity, including AGI and PI systems, can use to find information that is maximally different from information that the intelligent entity already possesses. Note that this approach is consistent with the general idea illustrated in Figure 48.

Figure 58 describes methods for validating the usefulness and safety of information gathered by a PSI or other intelligent entity, including AGI and PI systems, via simulation, review by humans, comparison to previous knowledge, and running safety checks.

7. Acquiring and Aligning Values with Human Values

Combining values and ethical information and resolving conflicts between different value systems is critical to the safe and ethical operation of advanced AI systems, including AI, AAAI, PSI, AGI, and PI systems. Although human involvement is preferable in circumstances where the stakes relate to human safety and ethics, as AI systems scale in speed and scope, it will become impractical to have human involvement or oversight over all critical decisions at some point. Therefore, means of training, customizing, educating, and influencing advanced AI systems, as well as means of delegating authority to them, will become increasingly important.

In the applicant's view, it is particularly important that the values and ethics driving advanced AI, including AGI and PI, should be representative of the human population. Flawed as it sometimes may appear, the democratic process of voting is one of the best ways to resolve ethical and values conflicts in a way that includes all humans. Therefore, methods of resolving value conflicts using voting and other democratic means are included, without limitation, as some of the means to combine values and resolve conflicts.

The following Figures detail a range of methods that intelligent systems can use to help ensure that the values and ethics of advanced AI, including AGI and PI, are aligned with human values and that humanity is safe in a world where advanced AI has outstripped human cognitive abilities.

Figure 59 illustrates an example of possible means by which the PSI, or any intelligent entity, including AGI and PI systems, can validate its information gathering activities, as does Figure 57.

Figure 60 describes a method that intelligent entities, including AI, AAAI, PSI, AGI, and PI systems, can use to achieve consensus on values and value-based behavior in various scenarios via a voting mechanism.

Figure 61 describes a method that intelligent entities, including AI, AAAI, PSI, AGI, and PI systems, can use to achieve consensus on values and value-based behavior via a weighted voting mechanism.

Figure 62 describes methods that intelligent entities, including AI, AAAI, PSI, AGI, and PI systems, can use to identify, analyze, weight, and optionally combine ethical or other information to reach a consensus.

Figure 63 describes a general method that intelligent entities, including AI, AAAI, PSI, AGI, and PI systems, can use to identify, elicit, and train on ethical information.

Figure 64 describes a method using (weighted) converging evidence that intelligent entities, including AI, AAAI, PSI, AGI, and PI systems, can use to determine ethical values, safety, or other information, and resolve conflicts by a prioritization process.

Figure 65 describes a method that intelligent entities, including humans, AI, AAAI, PSI, AGI, and PI systems, can use to delegate voting authority.

Figure 66 describes a reputational process that intelligent entities, including humans, AI, AAAI, PSI, AGI, and PI systems, can use to preserve information reflecting the views of a minority, including opinions on ethics, safety, or other issues.

Figure 67 describes a method that intelligent entities, including AI, AAAI, PSI, AGI, and PI systems, can use to make ethical, safety-related, and other decisions.

Figure 68 describes a method that advanced AI, including AI, AAAI, PSI, AGI, and PI systems, can use to learn, test, improve, and monitor safety and regulation-related rules and other information.

Figure 69 describes a method based on the Consequentialist Approach that intelligent entities, including humans, AI, AAAI, PSI, AGI, and PI systems, can use to make ethical, safety-related, and other decisions.

Figure 70 describes a method based on the Deontological Approach that intelligent entities, including humans, AI, AAAI, PSI, AGI, and PI systems, can use to make ethical, safety-related, and other decisions.

Figure 71 describes a method based on the Virtue Ethics Approach that intelligent entities, including humans, AI, AAAI, PSI, AGI, and PI systems, can use to make ethical, safety-related, and other decisions.

Figure 72 describes a method based on the Golden Mean Approach that intelligent entities, including humans, AI, AAAI, PSI, AGI, and PI systems, can use to make ethical, safety-related, and other decisions.

Figure 73 describes a method that trains an AI system, such as a foundational model or other trainable AI system, to be human-aligned and compliant with regulations.

Figure 74 describes a method to align a customized foundation model or other AI system (e.g., as described in Figure 73) with specific expertise or group ethics.

Figure 75 describes a method to form an AGI or PI that is aligned with human ethics and values, composed of other aligned intelligent entities, without limitation, including those described in Figure 73 and Figure 74.

As should be evident to AI researchers skilled in the art, the above methods can be adapted to apply at various levels to apply to individual intelligent entities (e.g. foundation models or AAAIs), networks of entities (e.g., AGI systems), networks of AGIs (e.g., PI systems) and networks of PI systems (e.g., IPI systems) with the over-arching goal of aligning such systems with human values, resolving conflicts between value systems, and/or delegating authority to various intelligent entities.

In aggregate, the above methods comprise multiple means to help ensure the various intelligent entities are aligned, and remain aligned, with human values, thus helping to ensure human safety. While the applicant, as a human, has a special interest in human safety, they also recognize and respect the rights and values of all sentient beings, which they believe does not constitute a conflict with human values about the most important and consequential ethical decisions.

8. Increasing Intelligence/Resources via Spot Market and Online Ad Technology

The applicant recognizes that any approach to creating safe advanced AI is unlikely to succeed unless the approach is also the fastest, most powerful, and profitable approach. Similarly, once created, an advanced AI will be unable to maintain its dominance and grow into a PI unless it has access to financial resources and the ability to generate profits that fuel the increase in its intelligence (e.g., via methods for catalyzing the growth of intelligence as outlined elsewhere in this application).

The applicant also recognizes that in the current business and research environment, those companies with the greatest chance of implementing AGI and PI successfully are those large technology companies that already have substantial capital, both financial and intellectual (specifically in terms of top AI research talent). These large technology companies, coincidentally, also derive a large share of their revenue and profits from selling online advertising. Thus, they have expertise, infrastructure, advanced AI systems, and access to many humans via their existing online advertising business efforts.

For these reasons, the applicant has invented technology enabling these large companies to leverage their existing online advertising infrastructure, combined with new inventive methods disclosed in the Figures below, so that they can develop the most advanced, profitable, and **safest** forms of AI, including AAAI, PSI, AGI, and PI systems, in the preferred implementation, these advanced forms of AI will not only be extremely profitable for these large technology companies but also align with human values based on the design of the systems and methods.

Of course, there is a risk in disclosing such potent new technology. Existing online advertising technology has not always served the interests of humanity well. However, the worst outcomes seem to occur when what is profitable conflicts with what is safe or good for humanity, or when the negative consequences of technology have been hard to see. Fortunately, business leaders recognize the potential danger of advanced AI. Even business leaders lacking AI expertise, such as Warren Buffett, acknowledge that tremendous risks exist alongside tremendous opportunities.

The current problem is not a lack of awareness of the dangers of AI, but rather a lack of knowledge of a safe path forward. In this situation, unwilling to halt development because of competitive pressures, leaders have largely settled for the most profitable path forward. Safety is given lip service. The standard response to calls for increased AI safety is for governments to regulate AI. After all, traditionally, regulation has been the solution to problems where what is profitable conflicts with what is safe. CEOs find it difficult to prioritize safety at the expense of profits when competitors are not forced to do the same thing. This situation could quickly prove disastrous for humanity, given that: a) the technology is developing much faster than regulators can act, and b) there is little hope of effective government regulation when the technologists themselves have no real idea how their systems work or how to make them safe.

The applicant believes that the only way out of this dilemma is to resolve the tension between profits and safety by inventing a path to advanced AI, AGI, and PI that is not only the safest, but also the fastest, most powerful, and most profitable. **No intelligent CEO would opt to create dangerous AI if they saw a way to create a safe AI that was more profitable.** Therefore, the Figures below describe methods that greatly increase online advertising profits while increasing the value, power, and safety of advanced AI systems.

Figure 76 generally describes the current technology for online advertising systems.

Figure 77 describes a method for implementing a spot market for the attention of expertise of intelligent entities (including but not limited to humans). This same market mechanism could be used for AGI or PI expertise.

Figure 78 describes a method for implementing the direct sale of attention or expertise from intelligent entities (including but not limited to humans). This same direct exchange mechanism could be used to buy and sell AGI or PI expertise.

Figure 79 describes a method for implementing an auction of attention or expertise from intelligent entities (including but not limited to humans). This same auction mechanism could be used to buy and sell AGI or PI expertise.

Figure 80 illustrates a system and methods for gathering expertise and cognitive work, including problem-solving work, from within an online ad unit and integrating it into an AGI problem-solving system. The approach can also integrate cognitive work into a PI system. This inventive approach enables monetizing online ads at a much higher rate since the value of expertise that can be trained and used by advanced AI many times greatly exceeds the value of humans making a single online purchase.

Figure 81 illustrates a method for improving online ad targeting, which is relevant to the system of Figure 80 by collecting and analyzing specific metrics, thus enhancing the capabilities of AGI or PI systems powered by the system of Figure 80.

Figure 82 illustrates a method for improving the effectiveness of the attention/expertise spot market of Figure 77 by collecting and analyzing specific metrics, thus enhancing the capabilities of AGI or PI systems.

9. Self-Awareness, Identities, and Conflict Resolution

An AGI or PI system will need a sense of self-awareness in order to operate effectively. The larger the scope of the system (e.g., especially in a global, planetary system), the more important this self-awareness becomes. AGI and PI will also have the ability to assume multiple identities to handle all the tasks that arise in parallel across the globe. Along with the identities comes the need for conflict resolution between conflicting identities and their associated priorities. The following Figures describe methods and attentional mechanisms that are useful for establishing and maintaining a sense of awareness and self-awareness in large, scalable, intelligent systems. The methods also cover many methods for resolving conflicts between identities, which is a challenge that becomes more complex and essential as the system scales from a single individual intelligence to a network of intelligences comprising AGI, to many networked AGI systems comprising PI.

Figure 83 illustrates conceptually the relationship between self-awareness (as a special case of), current awareness, and potential awareness. In the case of PI, potential awareness includes billions of sensors spread across the planet, all simultaneously feeding into awareness. Without systems and methods for directing attention and managing this input, the system would be incapable of functioning effectively as a PI.

Figure 84 illustrates the concept of multiple identities, using an example that is familiar to humans.

Figure 85 describes a general method for modelling awareness that intelligent entities, including AI, AAAI, PSI, AGI, and PI systems, can use.

Figure 86 describes the minimum required components that an intelligent entity, including AI, AAAI, PSI, AGI, and PI systems, must have to effectively shift attention and maintain awareness.

Figure 87 describes the process for setting parameters (dynamically) for working memory, which is a key component in the attentional system for an intelligent entity, including AI, AAAI, PSI, AGI, and PI systems.

Figure 88 describes a method for monitoring and updating awareness for an intelligent entity, including AI, AAAI, PSI, AGI, and PI systems.

Figure 89 describes an attentional interrupt system for an intelligent entity, including AI, AAAI, PSI, AGI, and PI systems. Attentional interrupts are critical for responding to dynamic conditions and new or unanticipated inputs that arise in awareness. From a safety perspective, attentional interrupts are essential for enabling the intelligent entity to respond to threats to human safety.

Figure 90 describes general methods that an intelligent entity, including humans, AI, AAAI, PSI, AGI, and PI systems, can use to gather input that can result in changing the entity's sense of identity. The methods are framed in terms that humans are familiar with, but analogs exist for non-human intelligent entities, as described in the text within the boxes of the figure.

Figure 91 describes some of the general methods that an intelligent entity, including AI, AAAI, PSI, AGI, and PI systems, might use to train the foundational model of the entity. The text of the Figure describes the model as a "Foundational Model", which would be tuned to create an AAAI. However, analogous methods can be used at the AGI and PI levels.

Figure 92 describes a specific version of a "Turing Test" that might be used with an intelligent entity, including AI, AAAI, PSI, AGI, and PI systems, to determine when the entity has been trained sufficiently. The text of the Figure describes a "Model", which could be trained or tuned to create an AAAI. However, analogous "Turing Test" methods can be used at the AGI and PI level. For example, when a PI is able to pass a version of the test that satisfies a statistically valid and representative sample of human beings on planet Earth, especially regarding its behavior across a wide range of safety, ethical, and values-based decision-making scenarios, the entity might be deemed sufficiently trained.

Figure 93 describes a method for arriving at a group identity (or identities) based on steps for combining many individual identities. This method is especially relevant for AGI and PI systems that must combine identities from many of the constituent intelligent entities, or networks of intelligent entities, from which they are comprised.

Figure 94 describes another method for arriving at a group identity (or identities) based on an explicit problem-solving goal to achieve the group identity and the use of the universal problem-solving framework and associated methods described elsewhere in this application. This method is especially relevant for AGI and PI systems that must combine identities from many of the constituent intelligent entities, or networks of intelligent entities, from which they are comprised.

Figure 95 describes a method for resolving conflicts between identities using a process that involves establishing a hierarchical identity structure with ethical/safety overrides. Such a method would be extremely useful, if not essential, for a PI attempting to handle millions or even billions of identities simultaneously.

Figure 96 describes a method that establishes, improves, and monitors behavioral protocols linked to identities. While useful for any intelligent entity, this method is especially useful, if not essential, for a PI attempting to handle millions or even billions of identities simultaneously.

Figure 97 describes methods for simulation and consequence prediction related to identities. Especially in high-stakes decisions and in situations where many identities are simultaneously active, as in the case of AGI or PI, the ability to simulate the effects of actions based on identities before actually taking those actions is essential. When a single human fails to plan and anticipate the consequences of an action, often that human suffers. If a PI fails to plan and anticipate via simulation, the consequences could be catastrophic for many humans and the entire planet.

Figure 98 describes a method for determining identity-based action using scenarios involving moral and other dilemmas. For reasons stated above, this method is especially important for large, complex systems such as AGIs and PIs.

Figure 99 describes a method for developing, refining, and evolving identities based on input from other intelligent entities. Again, for AGI and PI, where the stakes are high and the systems are complex, this method increases in importance.

Figure 100 describes a general process for reasoning ethically and predicting the consequences of actions in cases where conflicting identities suggest conflicting actions. The method is broadly applicable to intelligent entities, especially large, complex intelligences such as AGI and PI.

Figure 101 describes a method for resolving identity conflict using a process of hierarchical override with transparent justification that is subject to review and improvement. The method is broadly applicable to intelligent entities, especially large, complex intelligences such as AGI and PI.

Figure 102 describes a method for resolving identity conflict using an arbitration process with input from intelligent entities, ideally including humans. The method is broadly applicable to intelligent entities, especially large, complex intelligences such as AGI and PI.

Figure 103 describes a method for resolving identity conflict using negotiation and compromise. The method is broadly applicable to intelligent entities, especially large, complex intelligences such as AGI and PI.

Figure 104 describes a method for suspending and identifying (or identities) that may lead to destructive conflict or actions that threaten human safety. The method is broadly applicable to intelligent entities, especially large, complex intelligences such as AGI and PI. It is essential to have a safe design for advanced AI systems with self-awareness and identity-based decision-making.

10. Self-Extending Networks of AGI to Create PI

For an AGI network to expand into a network of AGI networks and ultimately into a PI that is global in scope, a scalable way to expand and extend the networks needs to exist. Further, in the preferred implementation, the networks of AAAIs that comprise AGI, the networks of AGIs that comprise PI, and ultimately the networks of PIs that will someday comprise IPI should all be self-extending. That is, without any direction from humans or other intelligent entities, a basic design characteristic of these networks should be that they seek to increase in scope, speed, and power (e.g., intelligence) so long as such extension and expansion remain aligned with human values and do not constitute a safety threat to humanity.

Since a common aspect of AAAIs, AGIs, and PIs is that the intelligent entities that comprise these systems, in the technology systems and methods disclosed here and in the cited PPAs and PCTs, all share the universal problem-solving framework and associated methods. This common aspect implies that for each intelligent entity, or network of such entities (e.g., at

AAAI level, AGI level, or PI level), a default goal of the intelligence can be to use available free resources that are not otherwise required to expand and extend the network. That is, if there is nothing else the intelligence or network of intelligence has to do, it should default to a goal of extending and expanding the scope, speed, and power of its intelligence network. This problem-solving goal can be realized in the same way that any other goal on the network is addressed, namely by solving the problem of “extending and expanding the network” using available resources on the network.

Just as individual humans often have a goal of increasing and expanding their individual intelligence, AAAIs should have a default goal of becoming more intelligent. Just as the human species acts as if it aims to increase the overall scope and collective intelligence of the human species, a network of AAAIs should have a default goal of increasing the number of AAAIs and other intelligent entities on the network. And just as life itself can be thought of as having a “goal” to increase the total amount of living species, so too AGIs (each composed of a network of AAAIs and intelligent entities) can have a goal of increasing the number and intelligence of AGIs in a network of AGIs comprising PI. Thus, a default goal of all artificial intelligence entities and networks can be to expand the scope, speed, and cognitive power, WITHIN the constraints of remaining aligned with human values and not endangering humans. Many specific methods can be developed to accomplish this goal via the general mechanism of problem-solving.

For example, Figure 105 illustrates the following exemplary method of an AGI network seeking to expand to create a PI, using the following process:

1. Multiple AAAIs, humans, and/or other intelligent entities participate in a network, comprising an AGI, as described in cited PPAs and PCTS.
2. Each time a client pays for problem-solving or other cognitive work on the AGI network, the system reserves a portion of the payment to cover operating costs, including a reserve allocated to expand the network.
3. Whenever some of the AGI networks are not engaged in solving problems for clients, the intelligent entities are recruited to achieve the goal of safely and ethically expanding and extending the AGI network itself, following the Universal Problem-Solving Framework of Figure 6; to wit:
    a. A default goal is set to expand the AGI network; the problem is activated.
    b. Run safety and ethics checks each time a goal or subgoal is set and also before each potential action is taken (e.g., as shown in Figure 107 and Figure 108) in the sub-steps below, while following this (exemplary) problem-solving process as long as spare capacity and resources exist to work on the problem:
        i. Intelligent entities are recruited to solve the problem.

ii. The intelligent entities represent the problem as one of achieving a series of sub-goals, via solving sub-problems, for example:
    1. Increase the intelligence of AAAIs on the existing AGI network
    2. Recruit additional human intelligences to the existing AGI network
    3. The more intelligent AAAIs and additional humans are used to determine the bottlenecks to the greater expansion of the network.
    4. Prioritize the bottlenecks such that the ones that lead to the most significant benefit in terms of network expansion are solved first.
    5. Apply means-ends analysis and other problem-solving techniques to solve each bottleneck and expand the network.
    6. Repeat from sub-step ii-4 until:
        a. Diminishing returns occur, in which case, assume that the easy progress on network expansion with given levels of intelligence has been achieved, and it is time to switch to increasing the intelligence of entities on the network as opposed to increasing the scope of the network, and revert to sub-step ii-1; or
        b. Spare resources are exhausted, and the default expansion problem is paused, awaiting additional resources to solve other client problems in (ii2).

The exemplary method in Figures 107 and 108 is only one of many variations of solution steps that might be followed to solve for the goal of expanding the network. Note that as all (successful or unsuccessful) solution attempts are recorded and analyzed (e.g., via methods described in Figures 8 and 9), the most effective ways to expand the network will be discovered over time. These can be stored, retrieved, and used preferentially to expand the network as efficiently and effectively as possible.

The same general process, as exemplified in Figures 107 and 108, can be used to expand a network of AAAIs to create a more powerful AGI and expand a network of AGIs to create a more powerful PI. When applied to AGIs, the main difference would be that the "intelligent entities" recruited in Step 1 of Figure 107 would not be AAAIs or humans, but rather AGIs (i.e., networks of humans and AAAIs) instead. Because the problem-solving process is universal and scalable, it can be executed by individual "smaller" intelligent entities like humans or AAAIs, as well as by AGIs that are comprised of many "smaller" intelligent entities. More powerful intelligences, such as AGIs, are just able to run the problem-solving process faster and with greater parallelism than a single entity or a smaller network could. The result is that the more powerful intelligences, such as AGIs and burgeoning networks of AGIs, will accelerate the pace of expansion very greatly compared to the capabilities of lesser intelligences. (This general concept, in which the same methods can be used in a collective intelligence network at different levels, was illustrated earlier in Figure 111.)

The main safety factor (for humans) is that the problem-solving architecture itself, which runs faster or slower depending on the level of intelligence of the problem-solving entity, always performs the safety and ethics checks at each step, no matter how fast it runs. That is, safety and ethics checks MUST be designed into the very problem-solving architecture itself, with provisions that they must not and cannot be overridden, to prevent a runaway expansion with potential unintentional or unsafe consequences for humanity. To the degree that these safety and ethics checks may require actual human oversight and monitoring, the speed of the expansion process will be "rate-limited", a good thing from a safety perspective.

It should be obvious to engineers, software developers, and AI researchers that just as one does not design an automobile or train without brakes or write software code to have an uninterruptible infinite loop, so too one should not omit safety/ethics checks and other means of human oversight and interruption from the problem-solving loop. To do so would be foolhardy and, in this case, could expose humanity to existential risk. Do not do it!

3.0 IMPLEMENTATION OF A PLANETARY INTELLIGENCE

The following sections describe how to implement a Planetary Intelligence (PI) using the inventive systems and methods described in Section 2.

First, Section 3.1 describes a PI implementation at the simplest, and highest level.

Next, Section 3.2 provides exemplary detail of major components for this high-level implementation and the relationships between components.

Next, Section 3.3 concisely details the potential inner workings of each major component by referencing the associated Figures previously explained in Section 2 that can individually, or in combination with other referenced methods, implement each major component. Section 3.3 also provides details of the various systems and methods in each of the categories of supporting systems in the same way, i.e., by concisely referencing methods in the previously disclosed Figures.

Finally, Section 3.4 provides a detailed example of one concrete PI implementation using an exemplary subset of the systems and methods referenced in Section 3.3.

3.1 High-level Description of PI System

Perhaps unsurprisingly, a global Super Intelligent AGI Network, a.k.a. “Planetary Intelligence” or “PI”, must be designed modularly to scale in intelligence and scope from much smaller and more limited components. As discussed above, PI comprises a network of Artificial General Intelligences (AGIs). Each AGI is a network of other intelligent entities, e.g., AAAIs, PSIs, and humans.

The AAAIs can be customized by various methods and incorporate various safety-related, resource-generating, intelligence-increasing, and learning systems and methods. When customized and personalized to the degree that the AAAIs can exhibit intelligence exceeding most humans at tasks, the AAAIs are called Personalized Super Intelligences (PSIs). However, for this disclosure, the applicant frequently uses the terms AI agent, AAAI, and PSI interchangeably. All are “artificial” or non-human intelligent entities. The networks of such individual intelligent entities can comprise an AGI, which is considered a more powerful intelligent entity. Similarly, networks of AGIs (i.e., “networks of networks” of intelligent entities) can also be thought of as very powerful intelligent entities, including ones that are SuperIntelligent and of global scope, namely PIs.

Humans, of course, are also intelligent entities. Networks of humans can exhibit “collective intelligence,” which is more powerful than the intelligence of a single human alone. The applicant’s prior work designing and implementing such human collective intelligence networks has proven that such networks can solve tough problems, such as getting an edge in the stock market, better than most professional expert humans. Two heads truly can be better than one. Imagine what two million can do!

The invention of creating AGI from a network of human and non-human intelligent entities that solve problems using a common and universal problem-solving framework, together with many ancillary systems and methods, has been disclosed at length in previously cited PPAs and PCTs, and in the Figures cited in Section 2.3 and reproduced as part of this application.

Since AGI was invented and designed in a modular way, and since problem-solving methods were developed that are truly universal and usable by any intelligent entity, at any scope, it makes sense that PI can be implemented or “assembled” using the same components as AGI, just at the “next level up.” The same systems and methods can be used as those developed for AAAIs and AGIs. Still, the intelligence unit moves from being an individual intelligent entity to a network of intelligent entities (for AGI) and then to a network of networks of individual intelligent entities (for PI).

The applicant has disclosed many possible combinations of the more than 100 systems and methods in this and previously cited PPAs and PCTs. Depending on the systems and methods used, one can produce different implementations of AAAI, PSI, AGI, and PI. A common feature of all collective intelligence systems, including CI systems of individual human and/or non-human intelligent entities (e.g., as in the case of AGI) and CI systems of networks of networks of intelligent entities (as in the case of PI) is the requirement for a universal problem-solving framework that enables rigorous communication and collaboration between intelligent entities.

Even within a single artificial intelligence entity, e.g., an AI agent, AAAI, or PSI, it is possible to have intelligent sub-components or sub-agents, as in the mixture of experts approaches used in some LLMs. In this case, the universal problem-solving architecture and all associated methods can also coordinate intelligence “within the brain,” so to speak, of an individual intelligent entity like an AAAI. While the applicant has explained this inventive use of the systems and methods disclosed in this and other cited PPAs and PCTs, because this application is primarily focused on PI, the examples, system architecture, and disclosure primarily concern network of intelligent entities and the methods are described in the context of facilitating the collective cognitive abilities of such networks, and networks of networks, resulting in PI.

Concerning the architecture of PI, Figure 109 simplifies this complexity by categorizing all the disclosed systems and methods as either **10**, representing modular components of a Global Super Intelligent AGI Network (PI), or **20**, as supportive systems and methods that can increase the PI system's safety, efficiency, and effectiveness in various respects.

3.2 Major Components and Categories of Supporting Systems and Methods for Exemplary PI Architecture

There are many potential implementations of a PI architecture that are possible using various combinations of the systems and methods disclosed above and in previously cited PPAs and PCTs. There are also multiple ways to structure the major modular components referred to by **10** in Figure 109, and the supportive systems and methods referred to by **20** in Figure 109. For exemplary purposes, one preferred implementation of an Architecture for Planetary Intelligence is illustrated in Figure 110.

Regarding Figure 110, **20** refers to the overall PI system.
In addition:

- **21** refers to systems and methods related to the online advertising technology inventions that enable better monetization of online ads, which can serve as a source of funding and financial resources for the system.

- **22** refers to the universal problem-solving framework, which serves as the means of rigorous communication for problem-solving and other cognition across intelligent entities.
- **23** refers to a network that pools or coordinates multiple intelligent entities using the universal problem-solving framework.
- **24** refers to the intelligence spot market, also referred to as “attention spot market” or “expertise spot market” in previously cited PPAs and PCTs.
- **25** refers to Advanced Autonomous AIs, AAAIs, sometimes referred to as AI agents, which can be customized and personalized.
- **26** refers to personalized AAAIs that have sufficient capability to exhibit super-intelligent (i.e., intelligence surpassing the average human) behavior in certain areas, and which are known as Personalized SuperIntelligences, or PSIs, and which can work together on a network.
- **27** refers to the Artificial General Intelligence that results from the combined intelligence of multiple intelligent entities, which can include AAAIs, AI agents, PSIs, and humans all working together.
- **28** refers to a network of AGIs that incorporate a sense of awareness, self-awareness, and multiple identities as described in previously cited PPAs and PCTs, and which together comprise a global superintelligent AGI network, or Planetary Intelligence.
- The arrow from Online Ad Tech (**21**) to Universal Problem-Solving Framework (**22**) represents the infusion of financial resources that can be associated with goals and problems on the CI network (**23**), and which can be used to reward problem solvers.
- Of course, financial resources can come from many sources, including clients will to pay for solutions to problems, but in this exemplary implementation, where methods for self-extending the network (**31**) apply, the Figure is illustrating how the system might use money generated by the Online Ad Tech invention to self-fund expansion of the network of AGIs and increase the scope of PI, e.g. by using funds to reward solving the self-extending problem and/or recruiting intelligent entities via the Intelligence Spot Market (**24**).
- The “**+**” between (**22**) and (**23**), and between (**23**) and (**24**), indicates that modules **22, 23**, and **24** together comprise important elements of the network that work together in combination.
    - This combination can be used internally:
        - to help AAAIs (as shown by the arrow to **25**), or
        - across multiple individual AAAIs and humans or other intelligent entities in a network of AAAIs (as shown by the two arrows to **26**), or
        - across networked PSIs, with other intelligent entities, to form AGI (as shown by the two arrows to **27**), or

- Across multiple AGIs, which themselves are networks of intelligent entities, to create a network of AGIs, which, if of sufficient global scope and capability, comprises a PI (as shown by the two arrows to **28**)

- Further, the PI (**20**) will be a self-aware PI if its components have awareness, self-awareness, and identity-related capabilities enabled, e.g., by the systems and methods for these capabilities that have been previously disclosed.
- The supportive systems and methods, generally referred to as **30**, comprise four categories of inventive systems and methods that increase the profitability and scope (**31**), ethics and safety (**32**), speed of intelligence growth (**33**), and learning and knowledge combination abilities (**34**) of the overall PI system (**20**).

3.3 Detailed Mapping of Inventive Systems and Methods to Exemplary PI Architecture

This Section provides a detailed mapping of inventive systems and methods to the exemplary PI architecture disclosed in Section 3.2. A concise way of accomplishing this mapping is to systematically examine each of the major components and categories of supporting systems and methods that were illustrated in Figure 110 and list the specific inventive methods (by citing Figures from Section 2.3) that would be used, in a preferred implementation, to implement the function of each major component or category. Multiple methods can be used individually or in combination with other methods for each component or category. Therefore, in the lists that follow, it is understood that the applicant intends that "any one or any combination of" methods in the list can apply for implementation purposes.

- With respect to Figure 110, **21**, the Online Ad Tech component, the relevant methods include, without limitation, those described in Figures 76, 80, and 81.

- With respect to Figure 110, **22**, the Universal Problem-Solving Framework component, the relevant methods include, without limitation, those described in Figures 3, 6, 11, 12, 14, 15, 16, 20, and 21.

- With respect to Figure 110, **23**, the Collective Intelligence Network component, the relevant methods include, without limitation, those described in Figures 2, 4, 5 (**items a-c, y, z,** and **a1 - m1**), 7, 11, and 12 (the AAAI.com network items in these Figures), and the problem-solving tree of Figure 17.

- With respect to Figure 110, **24**, the Intelligence Spot Market component, the relevant methods include, without limitation, those described in Figures 29, 46, 77, 78, 79, and 82.

- With respect to Figure 110, **25**, the AAAIs component, the relevant methods include, without limitation, those described in Figures 1, 5, 10, 13, 26, 27, 30, 31, 36, and 47.

- With respect to Figure 110, **26**, the Networked PSIs component, the relevant methods include, without limitation, those described in Figures 39, 40, 42, 44, 45, 74, and 75.

- With respect to Figure 110, **27**, the AGI component, the relevant methods include, without limitation, those described in Figures 28, 34, 35, 93, 94, 95, 96, 100, and 103.

- With respect to Figure 110, **28**, the Self-Aware Networked AGIs component, the relevant methods include, without limitation, those described in Figures 83, 84, 85, 86, 5 87, 88, and 89.

- With respect to Figure 110, **31**, the Self-Extending & Profitable supporting category of methods, the relevant methods include, without limitation, those described in Figures 107 and 108 and those already mentioned as relevant to (**20**).

- With respect to Figure 110, **32**, the Safety and Ethics Checks / Designed Features supporting category of methods, the relevant methods include, without limitation, those described in Figures 18, 19, 23, 41, 43, 59, 60, 61, 62, 63, 64, 65, 66, 67, 69, 70, 71, 72, 73, 101, 102, and 104.

- With respect to Figure 110, **33**, the Catalysts for Growth of Intelligence supporting category of methods, the relevant methods include, without limitation, those described in Figures 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, and 58.

- With respect to Figure 110, **34**, the Learning and Knowledge Modules/Combination supporting category of methods, the relevant methods include, without limitation, those described in Figures 8, 9, 22, 24, 25, 32, 37, 38, 68, 90, 91, 92, 97, 98, and 99.

3.4 Exemplary Specific PI Implementation Using a Subset of Systems and Methods

Imagine that you are the CEO of a major technology company, such as META, and you wish to implement a Planetary Intelligence system comprised of a network of powerful AGI systems. You might proceed as follows.

1. Concerning Figure 110 (**31**): You want the PI system to be self-funding and self-extending. Since META is already in the online advertising business, you decide to

implement the online ad technology inventions disclosed in PPA/PCT#8. You already have advertising technology systems and infrastructure similar to that described in Figure 76.

You instruct your development team to modify the "creative" of a large number of ad units so that they can work with the online problem-solving system described in this, as well as cited PPAs and PCTs.

Figure 80 describes the modifications that the team needs to make. You monetize the expertise of human experts targeted by the modified ad technology by selling the expertise and data gathered to other companies that are desperately seeking new sources of training data and expertise for their AI efforts. You can charge higher rates than a normal ad would generate for this type of valuable information that your modified ad tech is scooping up. You also use human experts' data to train your AI models, and the human experts themselves will help power the PI system.

2. With reference to Figure 110 (**22**), you need to funnel the human experts gathered in Step 1 to an online problem-solving system, where they will use the Universal Problem Solving Framework to collaborate with other intelligent entities, initially mainly other humans. Figure 6 describes the Universal Framework that the human experts will use.

   Figures 11 and 12 describe at a high level how they will solve problems, sequentially or in parallel.

   Figure 14 describes some of the methods, including matching human experts to problems, compensating the solvers, etc., that will be needed to implement the problem-solving system.

   Figure 16 identifies some key characteristics of problem solving, such as a goal-sub-goal hierarchy and means for recording solution attempts in an auditable record that will facilitate learning by the system.

   You don't want human experts to have to know anything about problem-solving theory or the universal architecture, so the system uses a Natural Language to Problem Solving Language Translator (as described in Figure 21) that enables the humans to work with regular natural language while the system translates everything into the rigorous problem solving framework that enables collaboration with other (human and non-human) intelligent entities.

3. With reference to Figure 110 (**23**), problem-solving takes place on a network, as described in Figure 2.

   Further operation of this network is described in detail in Figure 5 (**items a-c, y, z,** and **a1-m1**).

   The network tracks all problem-solving activity in a giant problem-solving tree structure, a.k.a. the "WorldThink Tree", which is described in Figure 17.

4. With reference to Figure 110 (**24**), for specific problems, META may purchase expertise, human attention, or attention from other intelligent entities via an intelligence spot market. META is an excellent position to actually run the spot market, since it has access to so many human users who might wish to participate. Since the value of human attention is linked to reputation metrics, META will likely administer an impartial reputational system as described in Figure 29 to help categorize the types of expertise offered on the spot market and facilitate price discovery. This same reputational system could also be used by the collective intelligence network in step 3, and reputations are also associated with AAAIs discussed below in step 5.

   Standardized tasks, as described in Figure 46, are important ways of benchmarking the performance of (human or non-human) intelligent entities, which can also affect pricing on the spot market.
   Figures 77, 78, and 79 provide additional details on the methods for implementing the spot market itself.

   Finally, Figure 82 describes how to optimize parameters relating to the operation of the spot market. Since the attention, expertise, and knowledge of intelligent entities are likely to become the most valuable commodity in the future, operating a spot market for them is both a lucrative source of revenue to fund PI and also a means of ensuring that the PI has access to the best intelligence available for any given task.

5. With reference to Figure 110 (**25**), META may choose to augment human experts on the networks with AI agents. Fortunately, META has developed some foundational AI agents itself (e.g., Llama 3) and has access to many more. However, to be most effective as intelligent entities on the network, these foundational models must be trained, tuned, and customized further. They need to become Advanced Autonomous AIs, or AAAIs. The roles that these AAAIs will play, how they can be customized, how they can use a problem-solving architecture, how they fit on the network, how they integrate, and how they improve are generally outlined in Figure 1.

A detailed implementation process for creating, training, and increasing the intelligence of these AAAIs is shown in Figure 5 in the "AAAI creation" section, while the top half of Figure 5 shows how they participate in the network to solve problems.

Another view of how they integrate, including with safety functions, is described in Figure 10.

To enhance their effectiveness, once trained, the AAAIs can be cloned as shown in Figure 13. Cloning allows many cloned AAAIs to participate in multiple problems in parallel.

Because META has access to so many human users and has extensive social media profile information on billions of users, this information can be used by META or by META's human users to customize and train customized versions of AAAIs, as described in Figure 26.

An important aspect of this customization and training is that as each AAAI is trained, it learns not only knowledge and expertise from a human user (and the user's profile) but also ethical and values information. The result is that META ends up owning, or having access to, billions of customized AAAIs, each reflecting slightly different values and ethical preferences. When aggregated, the preferences of the AAAIs constitute a representative and statistically valid sample of human values and ethics, which is essential for creating safe and ethical AGI.

Various methods, as illustrated in Figure 27, can be used to train and customize the AAAIs.

META may choose to use additional methods related to the development and deployment of AAAIs as described in Figures 30 and 31.

To help prevent the AAAIs from hallucinating when collaborating on problem-solving and to increase their reliability, META may want to implement the method described in Figure 36.

Finally, to optimize the performance of AAAIs that have been customized as PSIs, META can use the methods of Figure 47, producing multiple versions and optimizing parameters for problem solving.

6. With reference to Figure 110 (**26**), META will want to upgrade its AAAIs into PSIs by customizing and personalizing them, e.g., by using methods described in Figure 39.

PSIs will improve further as they work over long periods in a variety of scenarios (Figure 40), record and learn from solution attempts (Figure 42), and evolve via competition and selection (Figure 44).

The PSIs will work on a network as discussed in Figure 45, where, among other things, they can combine ethical preferences (Figure 74) and collectively, with other intelligent entities, comprise an AGI (Figure 75).

7. With reference to Figure 110 (**27**), Figure 28 describes generally how scalable and ethical AGI can arise from multiple intelligent entities that engage in problem-solving and refine their values.

   Figure 34 shows that one method for combining information, including ethical information, from multiple base LLMs, AAAIs, or PSIs is to experiment with combinations of the weight matrices in which these entities store their knowledge.

   Figure 35 expands upon this method.

   Figures 93, 94, 95, and 96 describe various methods for implementing identities that can aid in formulating a sense of self-awareness for these entities and also resolving potential conflicts resulting from different identities. The identity formation and conflict resolution process will aid the AGI in developing a robust sense of self-awareness.

   Further methods for dealing with identity conflicts, including methods of negotiation and compromise, are described in Figures 100 and 103. For safety and ethical reasons, META would likely want to implement these and other related methods before enabling full self-awareness at the AGI level.

8. With reference to Figure 110 (**28**), a PI should ultimately have a sense of global awareness, including mechanisms for tracking current awareness and locating its own sense of self-awareness within its current awareness, as illustrated in Figure 83.

   The PI (an even AGI, PSI, and AAAI systems) must be able to maintain multiple identities simultaneously and understand the relationships between these identities, as illustrated in Figure 84. While awareness and self-awareness could be implemented at the AI, AAAI, PSI, or AGI level, in this example, we are assuming that PI will be self-aware and networked.

To implement self-awareness, the PI must be equipped with attentional components as described in Figure 86.

Working memory parameters can adjust the scope of attention as described in Figure 87.

The PI must be capable of modelling awareness as described in Figure 85, and monitoring and updating its awareness as described in Figure 88.

For the awareness system to be efficient and safe, there must be attentional interrupts as described in Figure 89.

To reiterate (and as illustrated in the middle two boxes of Figure 111), AGI arises out of the combined collective intelligence of individual intelligent entities working together on a problem-solving network. PI arises out of the combined collective intelligence of AGIs, which have been networked together to extend the scope and capabilities of the PI system.

The PI becomes self-aware if methods for enabling attention, awareness, self-awareness, identity, and identity conflict resolution (as described in Step 8) are activated. While awareness can be activated at levels below PI, such awareness is likely essential at the PI level in order for the system to function effectively, given its global scope.

The steps 1-8 above provide a skeletal implementation example of PI, but many other methods are desirable, and in some cases, essential for effective and efficient operation. With reference to Figure 110, these additional methods have been categorized as supportive methods, referred to generally by the numeral (**30**). We now discuss some supportive methods and how they enhance the skeletal PI system just described.

With reference to Figure 110 (**31**), a general method for automatically extending the network of AGIs and thus increasing the scope of the PI would be beneficial, as described in Figures 107 and 108.

Further, PPA/PCT 8 describes the invention of new online ad technology and the intelligence, attention, or expertise spot market. These inventions can generate significant financial resources necessary to help the PI self-extend. One can imagine other problem-solving tasks, e.g., “develop a quantitative method for generating reliable and attractive returns with minimal risk in the US equity markets”, that could be posed to AAAI, PSI, AGI, or PI systems to generate further funding for expansion. Ultimately, there are direct correlations between the amount of intelligence that can be focused on a problem and the number of profits that can be generated, provided profitable problems are chosen. Given the projected superior intelligence of AGI and PI systems to any one or group of humans, it is reasonable to expect that such systems will ultimately self-fund their growth in intelligence.

Because PI is so powerful, the cited PPAs and PCTs disclose dozens of safety mechanisms, methods, and features to align PI with human values and ensure human safety. For example, with reference to Figure 110 (**32**), Figures 18 and 19 describe one of the most important and most scalable safety inventions, namely embedding of safety checks in the universal problem solving framework itself, such that no goal or sub-goal can be set, and no problem solving action taken, unless the goal or action passes a safety and ethics check.

This checking process, because it is embedded in the operation of the problem-solving process, scales with the speed of problem solving. That is, if PI can take one billion problem-solving steps in a millisecond, it must also perform one billion safety checks. Although human cognition cannot keep track of problem-solving at this speed, the safety checks are still being run extremely quickly.

Further, there are ethical dimensions to these checks, as described in Figure 23.

One of the most powerful ideas for keeping PI safe, even after it greatly outstrips human cognitive abilities, is the community safety method described in Figure 41. In this method, which applies to AAAIs, PSIs, AGIs, and PI, humans rely on a community of entities that are much smarter and more powerful than humans. Each entity was initially trained and customized based on human values. While it is possible, and perhaps likely, that a few of these super-intelligent entities may evolve and change their values such that they become malevolent towards humans, it is far less likely that most intelligent entities will evolve in this way. The method of Figure 41 halts the activities of the dangerous entities via oversight by human-aligned entities. As long as most of the intelligence remains human-aligned, this mechanism for advanced AI policing advanced AI should keep humans safe in the same way that Bitcoin is un-hackable, as long as the majority of participants in the system are honest.

Figure 42 describes how all problem-solving activities are recorded, e.g., using auditable and transparent blockchain technology. This record enables safety checks as described in Figure 43.

Simulations, with human validation, offer another avenue to increase safety as described in Figure 59.

Regarding ensuring ethical behavior of systems, multiple philosophically based approaches are possible, including but not limited to the Consequentialist Approach (Figure 69), the

Deontological Approach (Figure 70), the Virtue Ethics Approach (Figure 71), and the Golden Mean Method (Figure 72).

The methods of Figure 73 can be used to help ensure compliance of the AGI or PI with local and global regulations and ethical norms.

In situations where certain identities may lead to unsafe or unethical actions on the part of the AGI or PI, methods that allow human override, such as that described in Figure 101, can be used.

In less threatening cases of internal ethical conflict, arbitration with input from humans or other intelligent entities can help an AGI or PI resolve issues, as described in Figure 102.

The ability to halt actions, including suspending identities that lead to unsafe or unethical actions by an AGI or PI, is a critical safeguard, as described, for example, in Figure 104.

Many safety and ethics safeguards rely on the fact that the very architecture of the PI is composed of personalized and customized intelligent entities, each of which reflects the values of humans, and which, in aggregate form, is a representative and valid sample of human values. That is, the design of the AGI and PI networks is democratic in nature and thus offers the same level of ethics that humans are familiar with in democratic societies.

Figures 60 and 61 describe standard and weighted voting methods that help an AGI or PI integrate values democratically.

Methods for delegating voting authority, as described in Figure 65, are also helpful in this regard.

Figure 66 describes an innovative approach to using reputational concepts to preserve minority viewpoints and avoid “herd” decisions that might result in unethical or unsafe decisions by an AGI or PI.

AGI and PI systems can use methods based on converging evidence, as described in Figure 64, to help identify which human values apply in each situation.

Experiments, focus groups, interviews with humans, and other methods, as described in Figure 63, are also helpful.

Democratic voting methods to help AGI and PI systems adopt or combine value systems are describes extensively in a range of methods such is imperfect, but it avoids concentration of power, which history has shown to be a major threat to human well-being in cases where the values of the single powerful entity are not aligned with the welfare of the people.

With reference to Figure 110 (**33**), once META (in our example) develops a large number of AAAIs, which are customized and personalized, the capabilities of META's AGI network depend critically on the abilities and intelligence of these component AAAIs. Ideally, they are PSIs, capable of SuperIntelligent performance in at least some areas of cognitive activity. Many PSIs would enable a SuperIntelligent AGI. However, regardless of the level of intelligence exhibited by META's AGI or PI network, competitive advantage comes from increasing the intelligence of these entities as quickly as possible. The methods for catalyzing the growth of intelligence are relevant here.

For example, across all the component PSIs, and for the AGI and PI networks, the most useful new source of information for growing that intelligence should be identified as described in Figure 49.

Every source of potential new information should be categorized along multiple dimensions as described in Figure 50.

Algorithms related to the compressibility of the information could be applied to determine the information content of the new information sources, as described in Figure 51.

Parameters from Kaplan Information Theory (KIT) could be applied to further assess the usefulness of the information sources as described in Figure 52.

The distribution of tasks (and associated goals) that META's AGI of PI encounters could be calculated, and then the method described in Figures 53 and 54 could be applied to help prioritize new datasets that might most quickly improve the capabilities of the AGI or PI relative to its most frequent goals.

The AGI or PI could balance the cost of acquiring information against its usefulness as described in Figure 55.

Different versions of the PI, AGI, or of individual PSIs could be compared in terms of their effectiveness, and the methods described in Figures 56 and 57 could be further used to increase the intelligence of these entities.

Finally, heuristics are described in Figure 58, such as information that is maximally different from the entity's existing views, or seeking more recent information, or using principles such as converging evidence to increase reliability, can be employed.

With reference to Figure 110 (**34**), PI must continually learn and acquire new knowledge in order to increase its cognitive capabilities. One way that an AGI or PI system can learn is by storing and indexing successful solutions to problems on the network, via procedural learning methods as described in Figures 8 and 9.

Learning at the AAAI level also helps increase the cognitive capabilities of an AGI or PI by increasing the intelligence of the component entities. For example, Figure 22 describes training methods to help AI entities learn, customize, and gain specialized knowledge.

Acquiring new data from humans is often a prerequisite for further learning by an AGI or PI. In this regard, access to methods that extract such information, like the methods for automatic generation of questionnaires described in Figure 24, or the methods for identifying patterns in human behavior illustrated in Figure 25, can be helpful.

AGI and PIs may rapidly increase their knowledge by the direct combination of knowledge contained in the weight matrices of their constituent AAAIs, e.g., as described in Figures 32 and 37.

As described in Figure 38, combinations of knowledge from multiple entities can be accomplished by directly combining weight matrices or by combining training datasets and training constituent entities of an AGI or PI directly from the combined training data.

PIs can learn safety and regulation compliance information via multiple methods for training their constituent entities, some of which are described in Figure 68.

Since PIs will have many different identities, due to their vast scope, general methods for learning new identities or changing an entity's sense of identity can be helpful, as described in Figure 90.

META (in this example) would want its AGI or PI to attain a level of intelligence where it could interact with humans as capably as the most capable humans. Using methods such as the modified "Turing Test" described in Figure 92 can help evaluate when the PI system has achieved the desired level of cognitive performance.

The more complex an AGI or PI becomes, the more important simulation and scenario modelling become. Therefore, methods such as those described in Figures 97 and 98 can be especially helpful in maximizing the chances that AGI or PI acts in positive, human-aligned ways.

Finally, if PI is to be safe for humanity, it is important for PI to collaborate and continuously co-evolve with human society, as described in the methods of Figure 99.

If META's PI system incorporated all the above methods, in a preferred implementation, it would be much more capable, powerful, safe, and human-aligned than any existing AI system. Other combinations of methods cited in PPA/PCTs #1 - 10 are possible and may be desirable depending on the specific goals and purposes of the PI. However, the above exemplary preferred implementation represents a novel and useful initial design for PI.

Figure 112 is a diagrammatic representation of a computer system **100** that is utilizable or implementable with the user's device and/or any peripheral component of the present technology. *Note: Most information in Figure 12 is included in the following description, with the expanded details in the PCT.*

- The computer system **100** can be part of an example machine, which is an example of one or more of the computers referred to herein, and within which a set of instructions for causing the machine to perform any one or more of the methodologies discussed herein may be executed. In various example embodiments, the machine operates as a standalone device or may be connected (e.g., networked) to other machines. In a networked deployment, the machine may operate in the capacity of a server or a client machine in a server-client network environment, or as a peer machine in a peer-to-peer (or distributed) network environment. The machine may be a personal computer (PC), a tablet PC, a set-top box (STB), a personal digital assistant (PDA), a cellular telephone, a web appliance, a network router, switch, or bridge, or any machine capable of executing a set of instructions (sequential or otherwise) that specify actions to be taken by that machine.

- The computerized system **100** can include one or more processors **102**, storage devices **106**, communication devices, and software components or instructions **104** to provide a platform for users to interact with and train/tune the LLMs. The computing capabilities may be standalone or may be cloud-based. They may include cloud-based AI development platforms that seamlessly offer "AI as a service," including hardware and software components.

- The system also supports the ability for users to provide new data, or data that is unique to them, for the LLMs to learn from. The processors **102** may be one or more CPUs, GPUs, chips specialized for ML, microprocessors, application processors, embedded processors, field-programmable gate arrays (FPGAs), or other hardware components capable of executing computer programs. The processors may be in communication with one another and/or with other components of the system. Further, any one of or any combination of the components of the system **100** can communicate with each other via a bus **134**.

- The storage devices **106** may include one or more hard drives, solid-state drives, optical storage devices, or other storage components. The storage devices may store the data used to train/tune the LLMs and other data associated with the system, such as user accounts, system settings, and other data.

- The communication devices may include one or more cellular modems **108**, Wi-Fi cards **110**, Bluetooth modules **112**, Network Interface Device **114**, or other components that enable the system to communicate with other systems, such as user devices, over a network or the internet.

- The communication devices may enable the system to communicate with others over a wireless or wired connection **116**.

- The software components may include computer programs that provide a platform for users to interact with and train/tune the LLMs. The software components may also include computer programs for collecting, storing, and processing data used to train and/or tune the LLMs. These software components may also include computer programs that provide a user interface for users to interact with the system.

- Without limitation, the user interface **118** may include natural language interfaces, textual interfaces, and chatbot-type interfaces, a web-based user interface, a mobile application, an augmented reality application, a metaverse application, or other applications that allow users to interact with the system. The user interface may include features that allow users to select the data they want to use to train/tune the LLMs, as well as features that allow users to interact with and monitor the progress of the LLMs.

- The system may also include one or more datacenters, databases or data sources, including without limitation, vector databases, centralized databases, and distributed databases, for storing the data that is used to train/tune the LLMs, as well as other data associated with the system, such as user accounts, system settings, and other data. The databases may be hosted on the system or another system, including cloud-based systems.

- The system may also include one or more authentication systems to verify the identity of users who use the system and provide secure access to the system. The authentication systems may include biometric authentication systems **122**, such as facial recognition or fingerprint recognition systems, as well as other authentication systems, such as password-based authentication systems.

- The system may also include one or more security systems to protect the system from unauthorized access and the data that is stored on the system. The security systems may include firewalls, encryption systems, access control systems, single and multi-factor authentication systems, and other security systems.

- The system may also include one or more analytics systems for collecting and analyzing data associated with the system and/or the LLMs. The analytics systems may include machine learning algorithms and other algorithms for analyzing the data associated with the system and/or the LLMs.

- Data visualization methods, including the use of problem trees and other representations, and data structures; use of statistical outputs, tables, graphs, text, speech, video, image and graphical outputs may be used for one way or two-way communication between users and the system and between multiple (human or AI) agents or LLMs using the system to interact with each other in large or small groups.

- The system may also include one or more monitoring systems to monitor the performance of the system and/or the LLMs. The monitoring systems may include systems for monitoring the performance of the system, such as system uptime, and systems for monitoring the performance of the LLMs, such as accuracy, speed, ethical compliance, reputation metrics, quality metrics, and other metrics as discussed above or as are known in the art.

- The system may include one or more of the architectures described above that enable one or more human or AI Agents or LLMs to engage in a variety of intellectual tasks, including, without limitation, simple and complex and multi-step problem-solving behavior with the system having all the functionality and features previously described.

- The system may also include one or more feedback systems to allow users to provide feedback on the system and/or the LLMs. The feedback systems may include systems for allowing users to submit feedback on the system, such as bug reports, and systems for allowing users to submit feedback on the LLMs, such as suggestions for improving the accuracy or speed of the model.

- The system may also include one or more management systems for managing the system and/or the LLMs. The management systems may include systems for managing the system, such as systems for managing the users and user accounts, and systems for managing the LLMs, such as systems for managing the data used to train and/or tune the model.

- The system may also include one or more payment systems allowing users to pay for the use of the system and/or the LLMs. The payment systems may include systems for processing payments, such as credit card processing systems, and systems for managing payments, such as subscription management systems.

- The system may also include one or more other components, such as support systems, reporting systems, and other components necessary for providing users a platform to interact with and train/tune the LLMs.

- The computerized system of the present technology enables users to interact with and train/tune LLMs based on data that is unique to the users. The components of the system described herein provide the necessary hardware and software components to enable users to do so.

- Further, while only a single machine is illustrated, the term “machine” shall also be taken to include any collection of machines that individually or jointly execute a set (or multiple sets) of instructions to perform any one or more of the methodologies discussed herein.

- The computer system **100** may further include or be in operable communication with a video display **120** (e.g., a liquid crystal display (LCD), touch-sensitive display), input and/or output device(s) **130** (e.g., a keyboard, keypad, touchpad, touch display, buttons, sonic, sensorial, etc.), a cursor control device **132** (e.g., a mouse), a drive unit (also referred to as disk drive unit), and a signal generation device **128** (e.g., a speaker). The drive unit **124** can include a computer or machine-readable medium **126** on which one or more sets of instructions and data structures (e.g., instructions **104**) are stored, embodying, or utilizing any one or more of the methodologies or functions described herein. Instructions **104** may also reside, completely or at least partially, within memory **106** and/or processors **102** during the execution of the computer system **100**. The memory **106** and/or the processors **102** may also constitute machine-readable media.

- Still further, the computer system **100** can be in operable association or communication with any type of multi-modal input and/or output **130** that addresses the human senses, as well as I/O technology that extends beyond the range of normal human perception, such as the ability to process information invisible to humans, for example, but not limited to, X-rays and information outside of the typical bandwidths of human perception, but not outside of AI perception using tools. Additionally, the I/O technology can include very fast perceptions that are too fast for humans to perceive but which an AI entity could perceive, and very slow or faint perceptions (e.g., tiny seismic shifts occurring over years) that humans cannot perceive but which AIs could. Since any intelligent entity can be part of the present technology system, it can be appreciated that any type of I/O that humans and AIs with much broader perceptual capabilities than humans can utilize with the system **100**.

- Instructions **104** may further be transmitted or received over a network via the network interface device **114** utilizing any one of several well-known transfer protocols (e.g., HTTP). While the machine-readable medium is shown in an example embodiment to be a single medium, the term “computer-readable medium” should be taken to include a single medium or multiple media (e.g., a centralized or distributed database, vector databases, and/or associated caches and servers) that store one or more sets of instructions. The term “computer-readable medium” shall also be taken to include any medium that is capable of storing, encoding, or carrying a set of instructions for execution by the machine and that causes the machine to perform any one of or more of the methodologies of the present application, or that is capable of storing, encoding, or carrying data structures utilized by or associated with such a set of instructions. The term “computer-readable medium” shall accordingly be taken to include, but not be limited to, solid-state memories, optical and magnetic media, and carrier wave signals. Such media may also include, without limitation, hard disks, floppy disks, flash memory cards, digital video disks, random access memory (RAM), read-only memory (ROM), and the like. The example embodiments described herein may be implemented in an operating environment comprising software installed on a computer, in hardware, or in a combination of software and hardware.

- An example machine system of the present technology, including the computer system **100** in combination and/or operational use with components of the present technology. In the exemplary, any or all of the above-described components can include a processor **102**, memory **106**, a network interface device **114**, a display **120**, an input device(s) **130**, **132**, and/or a drive unit **124**.

4.0 CONCLUDING REMARKS

While tremendous opportunities exist for advanced forms of AI, including AAAIs, PSIs, AGI, and PI to benefit humanity, advanced AI also poses existential risks. To a degree greater than any technology previously invented, AI has the potential both to eliminate all forms of human poverty and material suffering and all forms of human life.

It is axiomatic that safety must be designed into complex systems. Safety and ethics cannot be “tested in” after the fact. Therefore, it is the responsibility of those inventing AI systems, especially AGI and PI systems, to ensure that safe and ethical, human-aligned behavior is a natural consequence of how the systems operate, not an afterthought achieved at the cost of reduced efficiency or effectiveness.

In this and the preceding nine applications, the applicant has disclosed a general approach to designing intelligent systems in which broader and more capable forms of intelligence emerge from the collective intelligence of lesser intelligent entities that work together using a common cognitive framework. This approach can produce customizable super-intelligent AI agents. It can produce AGI from the collective efforts of such agents working together on a network. It can produce Planetary Intelligence (PI) from the collective efforts of AGIs working together in a network of networks.

Many dozens of specific inventive methods have been disclosed that can be used to implement the approach. Most importantly, human intelligence is integral to the design of the systems. Humans serve the dual purpose of bootstrapping the cognitive abilities of AGI and PI and aligning these intelligences with human values and ethics.

It is true that even if all the safety-related methods are perfectly implemented in an AGI or PI system, some specific technical risks remain, as discussed in Section 1.4. However, the most significant risk for humanity is likely not that a properly designed AGI or PI system will go rogue and adopt a wholly new and different set of values and ethics that lead to human extinction. Instead, the greater risk, in the applicant's view, is that humans will intentionally or unintentionally model negative or destructive values that these sophisticated forms of AI will adopt and amplify.

The AGI and PI systems in this series of inventions use empirical methods to help advanced AI determine what constitutes safe and ethical behavior by observing human behavior and incorporating the value systems of millions of individual AIs that reflect a broad and statistically valid sample of humans. Despite media focus on war, terror, and generally the worst of humanity's behavior, statistically speaking, the vast preponderance of human behavior is prosocial and positive. AI systems with the level of intelligence envisioned in these patents should be able to accurately distinguish human behavior as it is, as opposed to how it is sensationalized in the media.

Of course, if our behavior deteriorates, if most of our behavior becomes excessively motivated by fear, greed, hatred, or negative values, then all bets are off. Just as children look to their parents and peers for their initial values, advanced AI will look to humanity, at least initially, for answers to questions related to morals and purpose. Since there is no purely logical way to determine what is right or wrong, it will be up to us to teach our AI children well. We must succeed or perish.

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

FIGURE 1: FIVE FUNCTIONS SPANNING LEVELS OF INTELLIGENCE

A flow chart illustrating an embodiment of the five subsystems utilizable in the present technology that operate across levels of intelligence and that are also part of the exemplary implementation of an AGI, or PI (comprised of a network of AGIs).

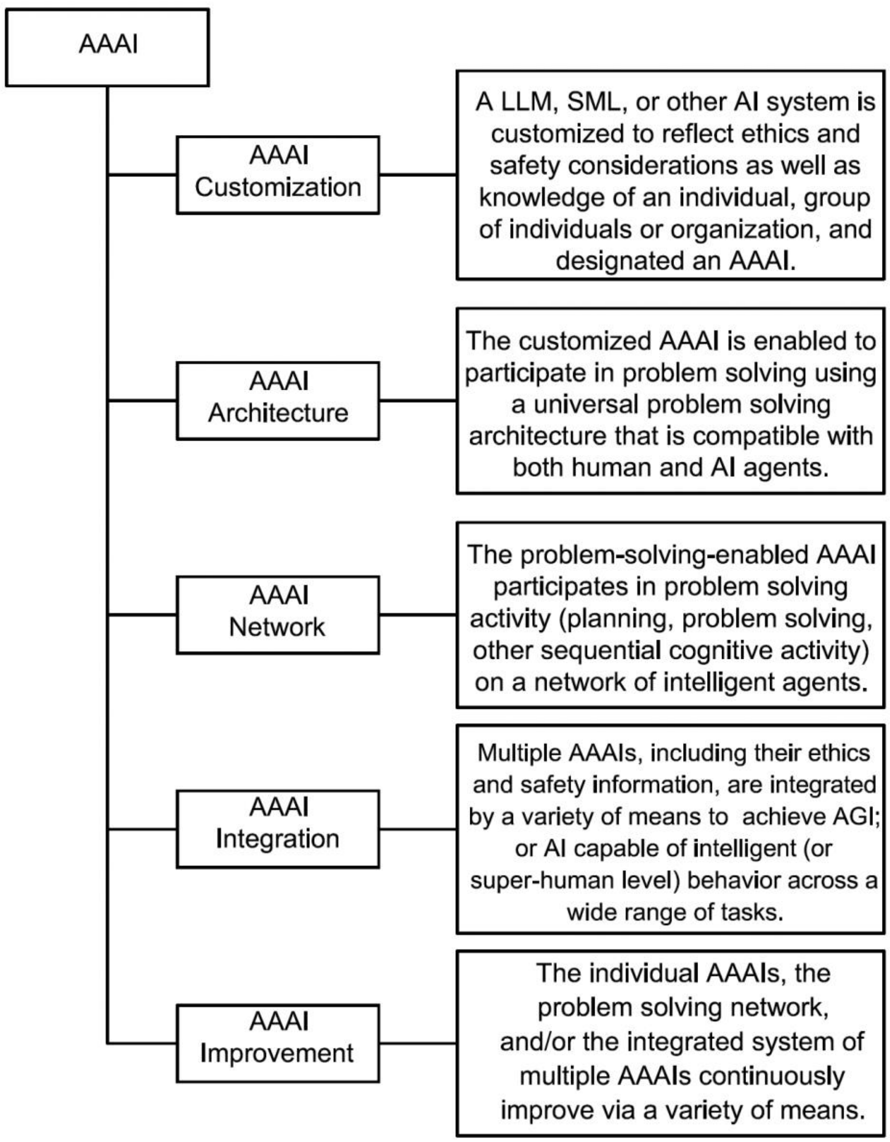

FIG. 1

SUPERINTELLIGENCE
AN iQ COMPANY

FIGURE 2: HUMANS AND AAAIS COLLABORATE ON A NETWORK TO FORM AN AGI

A block diagram illustrating an exemplary process of the overall process utilizable with the present technology and implementable in a web-based system where multiple individual intelligent entities (e.g., humans and AAAIs) collaborate on a network to comprise an AGI.

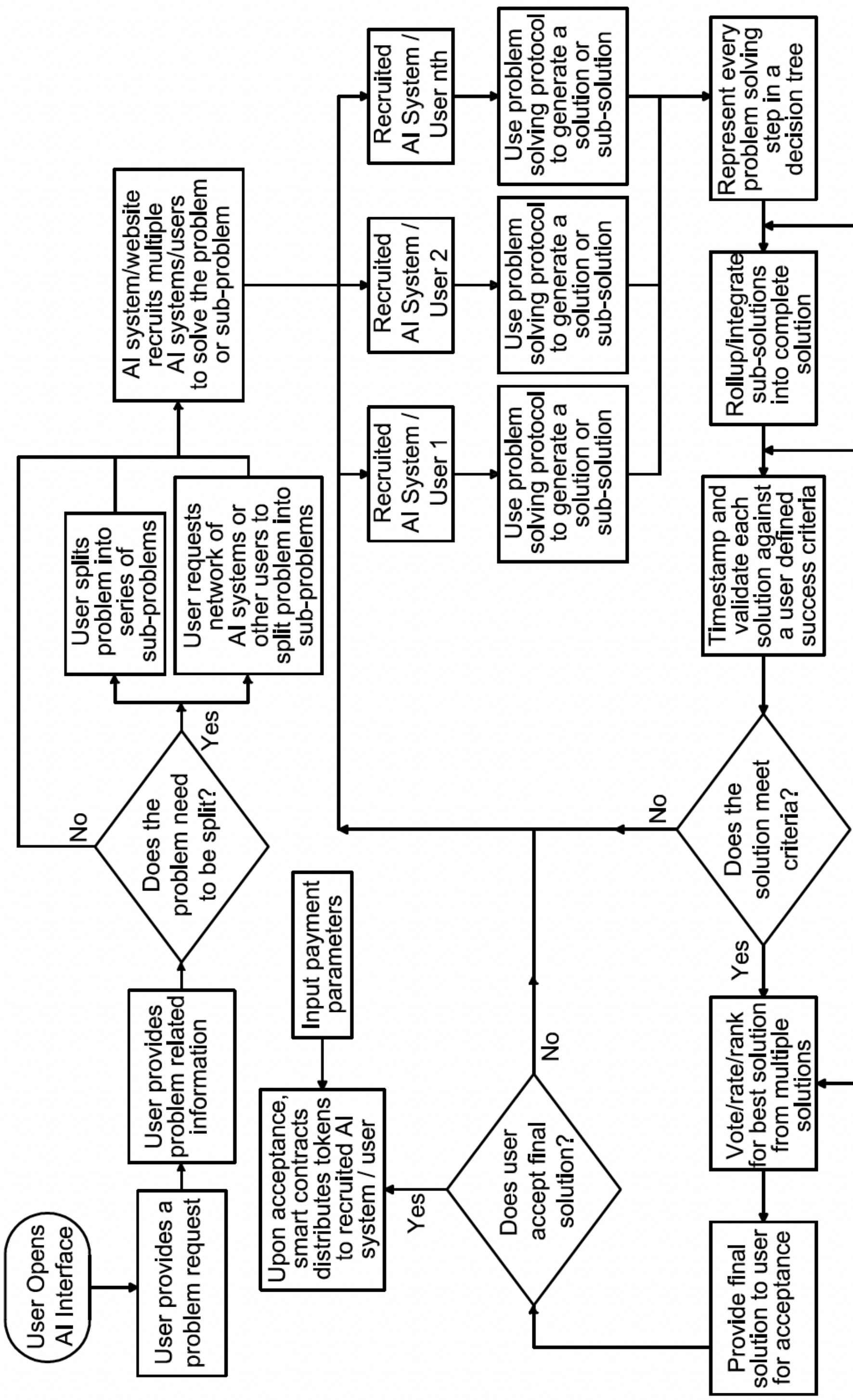

FIG. 2

FIGURE 3: PROBLEM SOLVING AS A DECISION TREE

A flow chart illustrating an example of how problem-solving can be represented as a decision tree using the current technology.

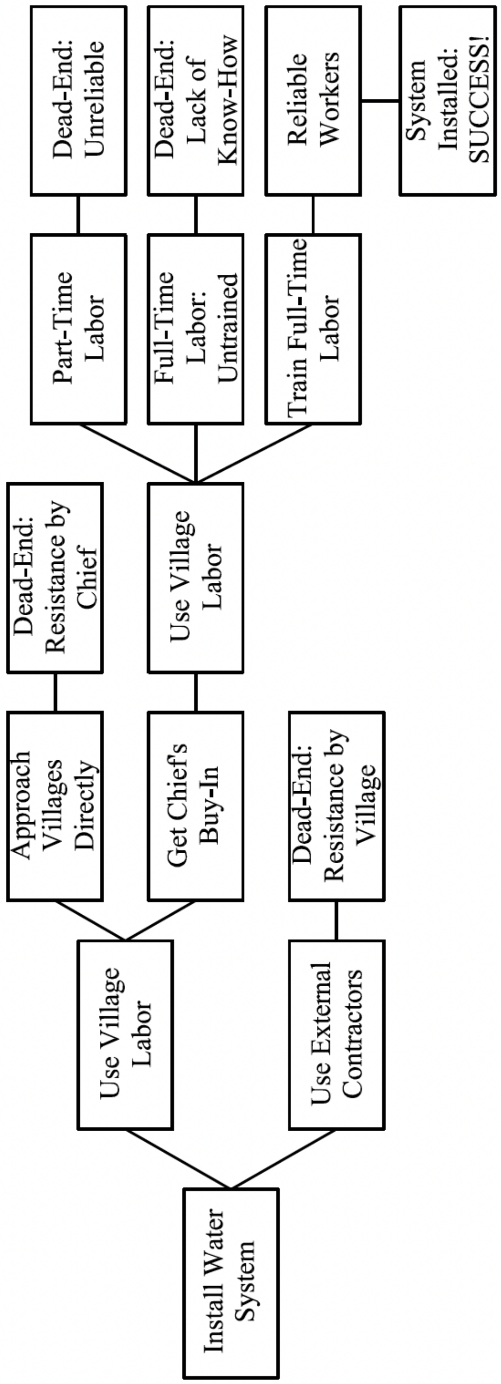

FIG. 3

FIGURE 4: WORLDTHINK PROTOCOL ENABLES AAAIS TO FORM AGI

A diagram illustrating how the underlying WorldThink protocol (including the universal problem-solving architecture, the problem tree, and all related components) supports the development of AAAI entities that are expert in various domains, and how those AAAI entities together support a larger and more powerful AGI formed from their collective intelligence.

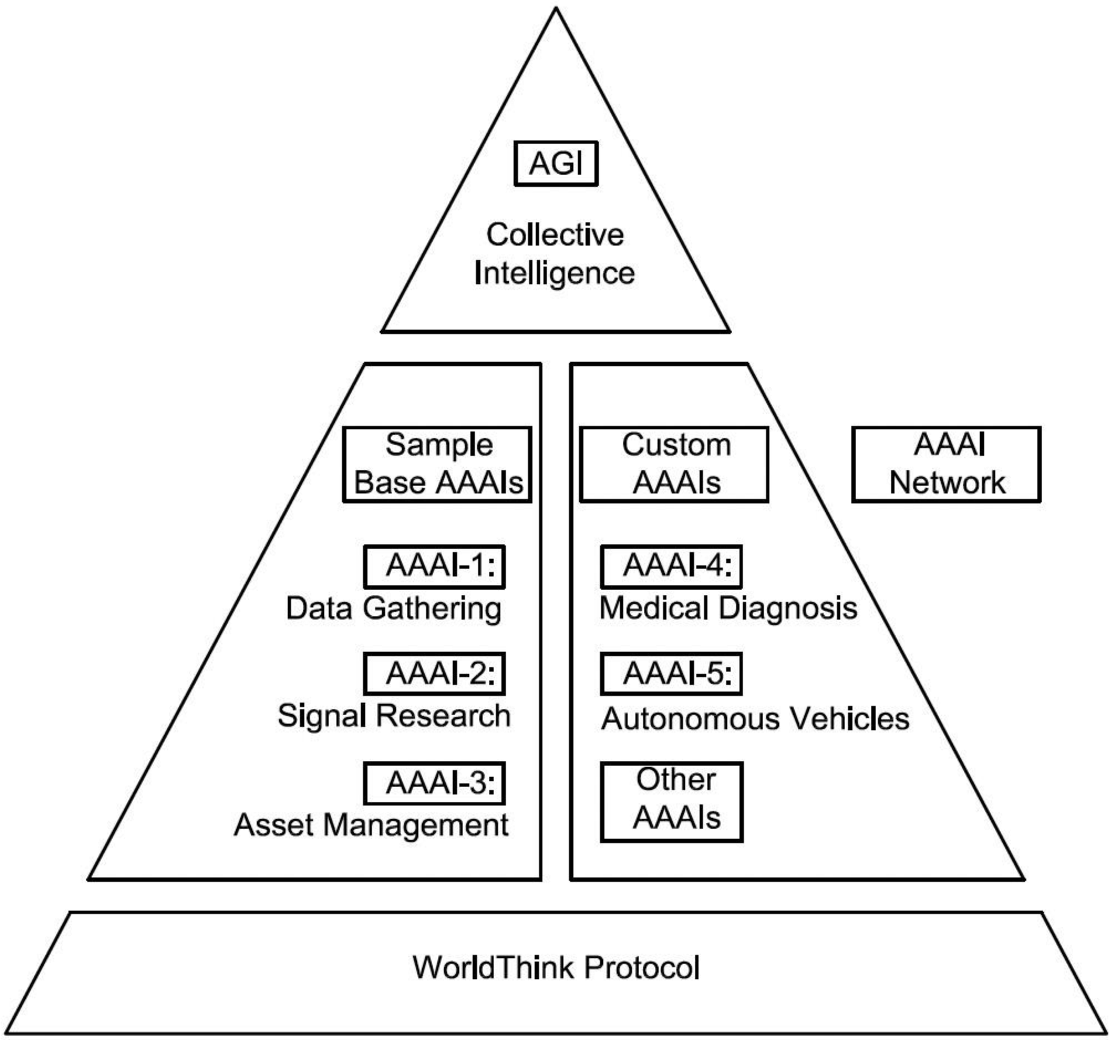

FIG. 4

FIGURE 5: TRAINING AND NETWORKING AAAIS TO SOLVE PROBLEMS

A detailed illustration of one implementation of the various methods and components needed to train and increase the intelligence of AAAIs and then put them to work on a network with other intelligent entities (including, optionally, humans) to solve problems using the tree structure and different problem solving methods of the current technology, including means for overall system learning and compensation of the solvers.

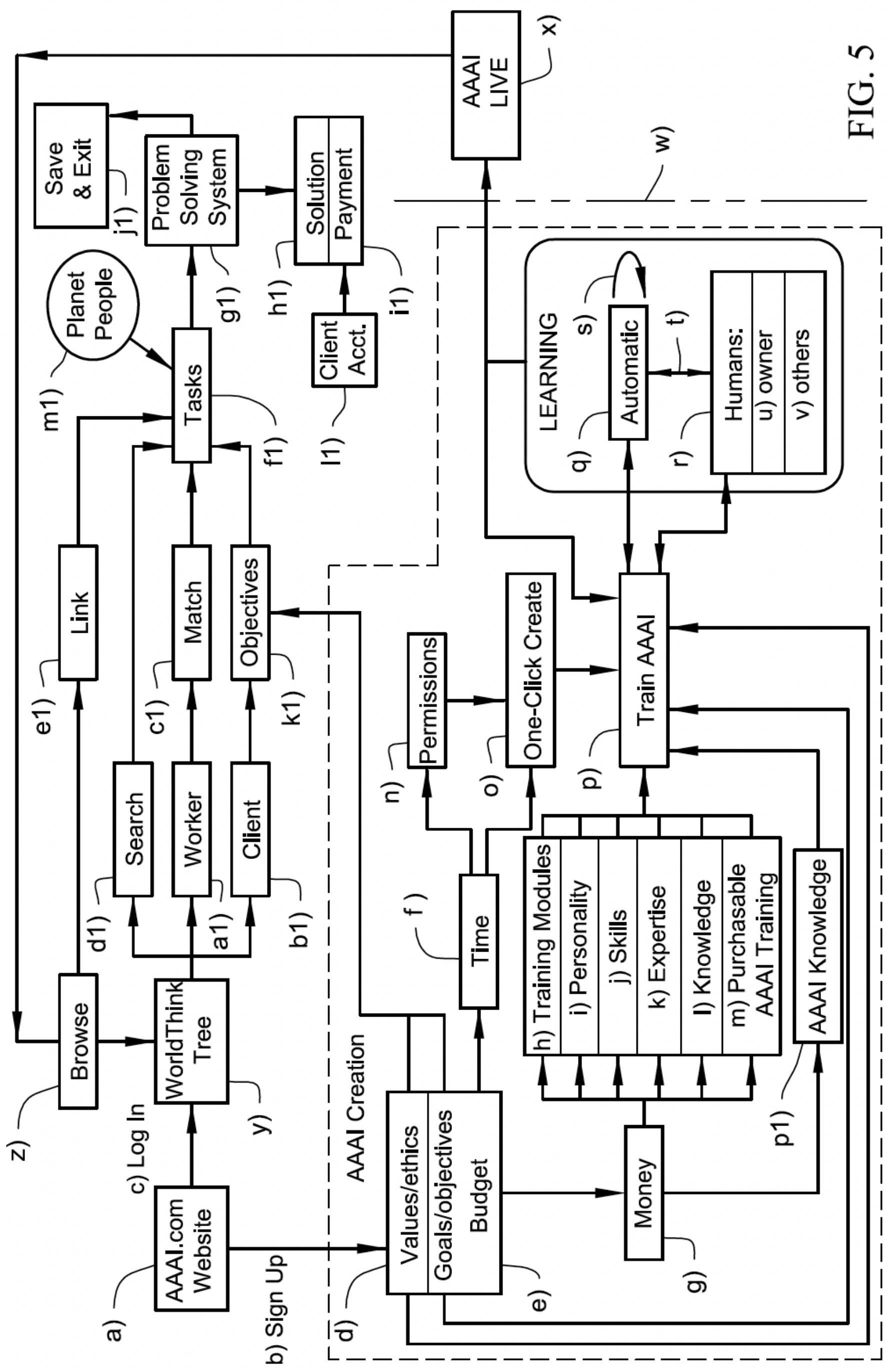

FIG. 5

FIGURE 6: UNIVERSAL PROBLEM-SOLVING ARCHITECTURE AND FRAMEWORK

Illustrates the scalable problem-solving framework of the present technology and the Universal Problem-Solving Architecture, a feature of the AAAI architecture box of Figure 1.

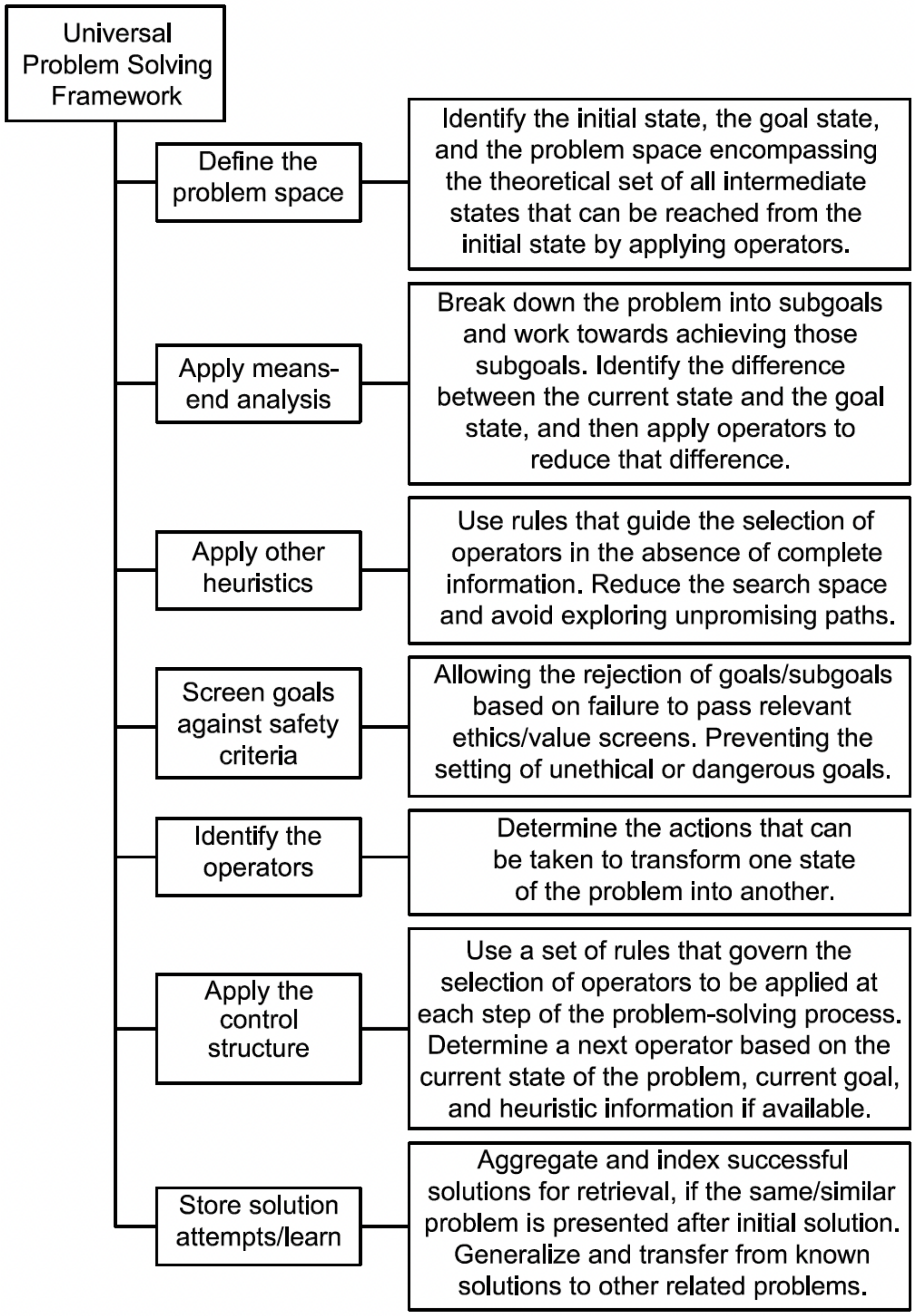

FIG. 6

FIGURE 7: PROBLEM-SOLVING ACROSS COLLECTIVE INTELLIGENCE LEVELS

A flow chart illustrating an exemplary problem-solving process utilizing a typical cognitive architecture implemented in a collective intelligence network, as referenced by the AAAI Network box of Figure 1.

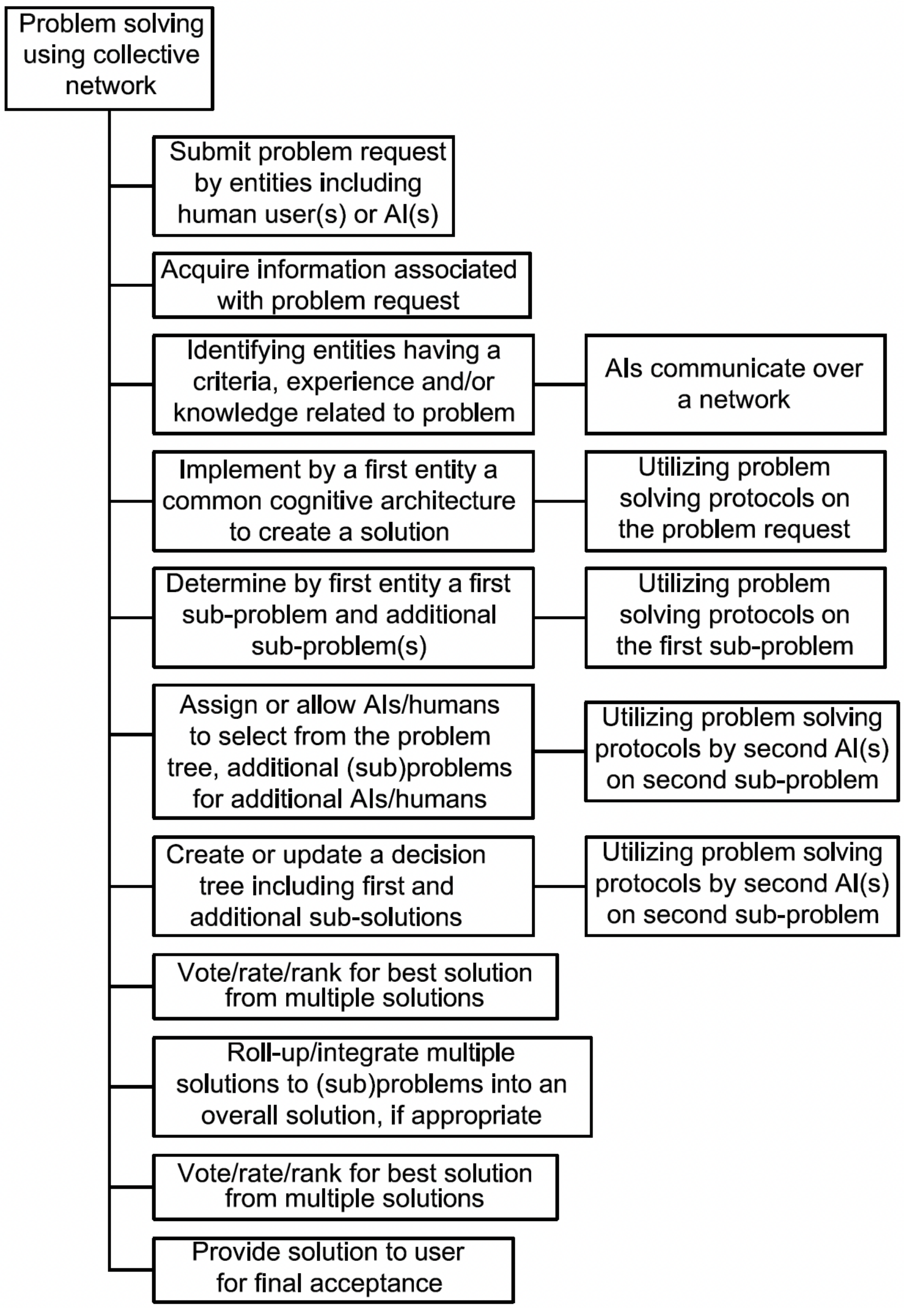

FIG. 7

FIGURE 8: PROCEDURAL AND SOLUTION LEARNING IN AAAIS AND AGI NETWORK

A flow chart illustrating an exemplary embodiment of the procedural learning process of the present technology, which is a way that AAAIs improve (as referenced in the last box of Figure 1).

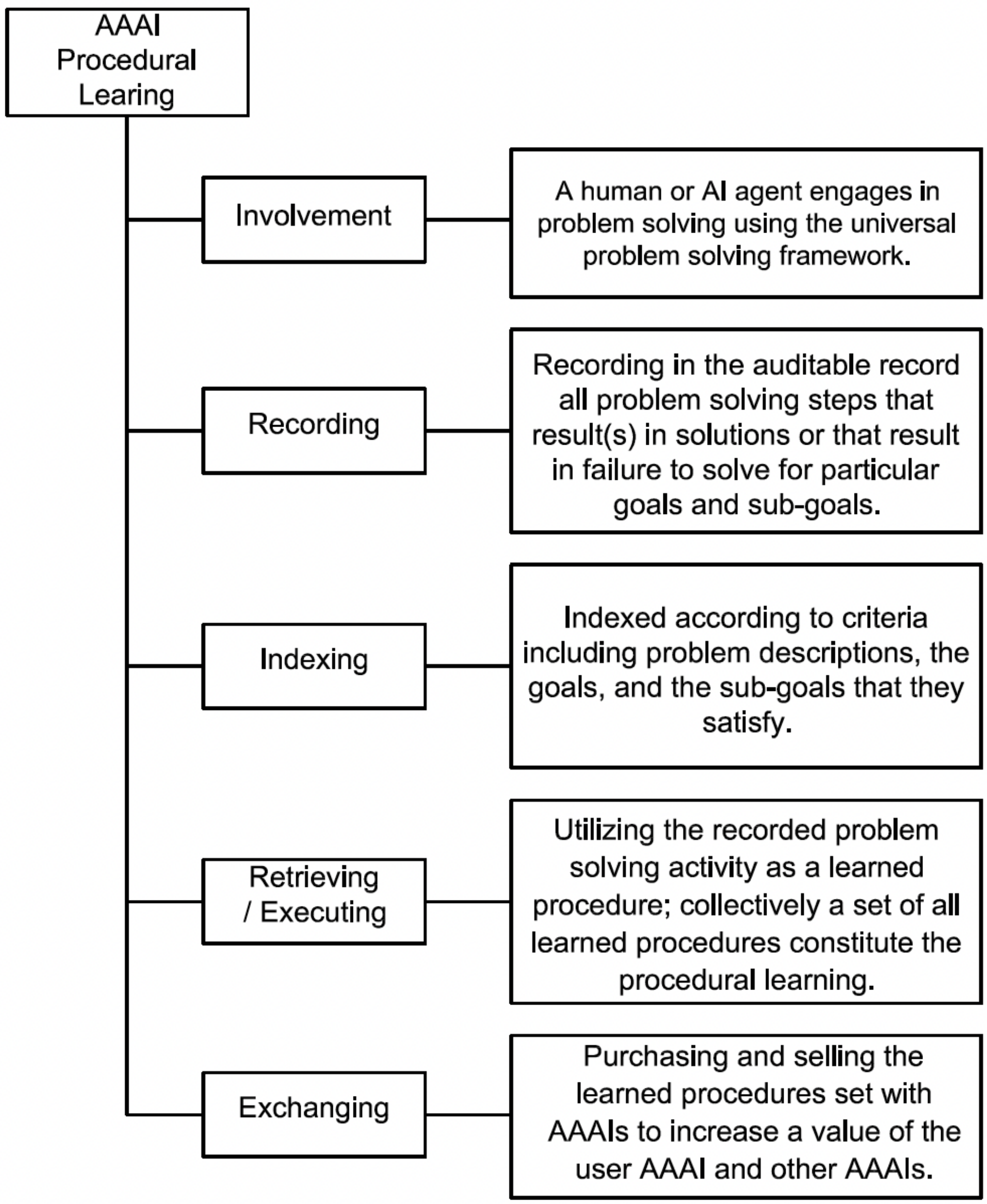

FIG. 8

FIGURE 9: LEARNING PROCESSES OF AAAIS AND AGI

A flow chart illustrating an exemplary embodiment of the solution learning subsystem or process.

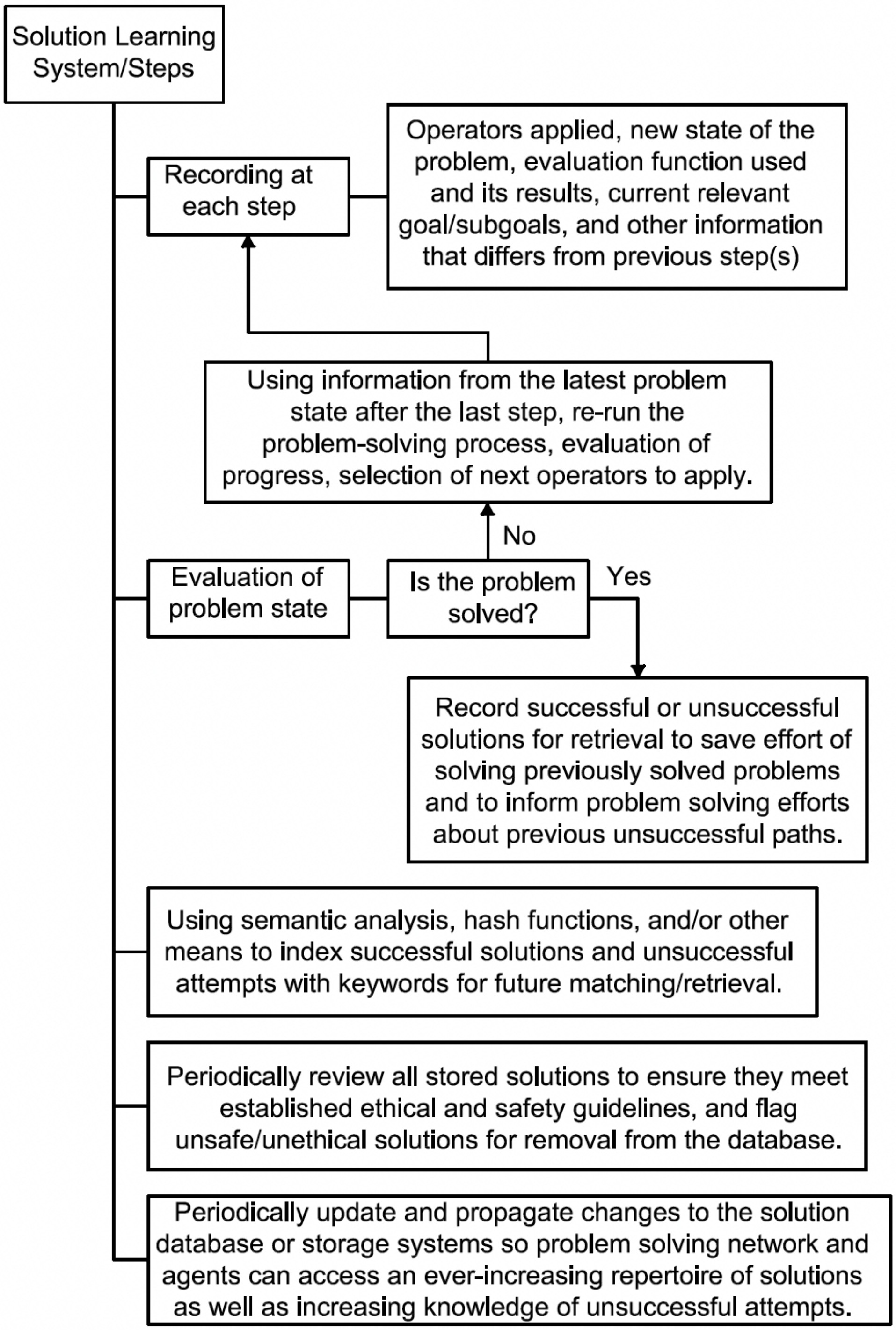

FIG. 9

FIGURE 10: DETAILS AND RELATIONSHIPS OF FIGURE 1 SUBSYSTEMS

A flow chart illustrating an exemplary overall process of the present technology.

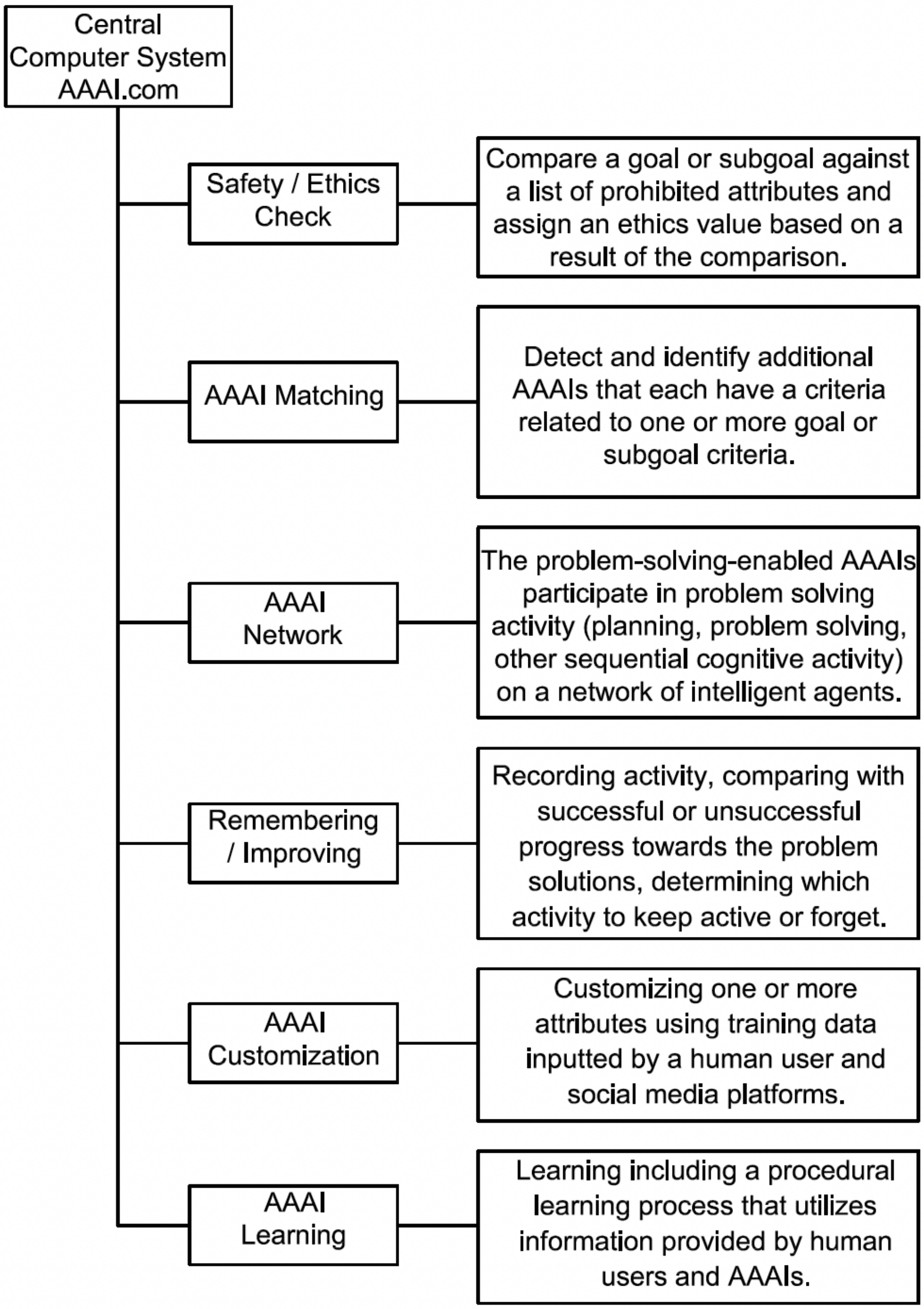

FIG. 10

FIGURE 11: SERIAL PROBLEM-SOLVING VIA WORLDTHINK PROTOCOL

A flow chart illustrating some of the basic problem-solving functionality supported by the WorldThink protocol, utilizable with the AAAI system and method of the present technology.

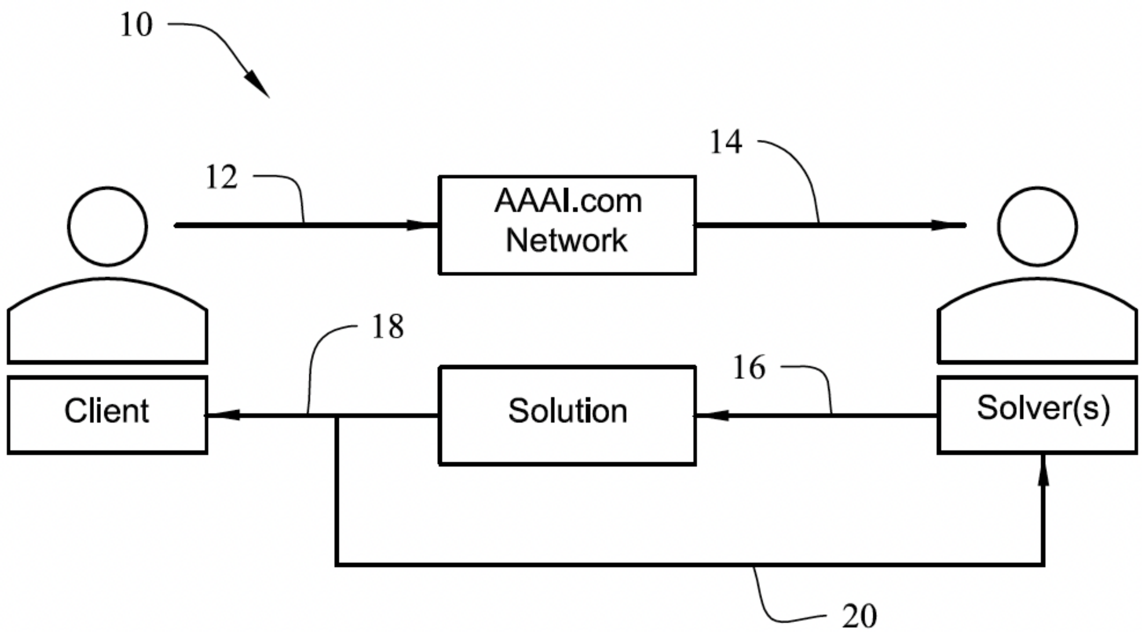

FIG. 11

FIGURE 12: PARALLEL PROBLEM-SOLVING VIA WORLDTHINK PROTOCOL

A flow chart illustrating some of the basic problem-solving functionality supported by the WorldThink protocol, utilizing two problem solvers collaborating to solve a client problem.

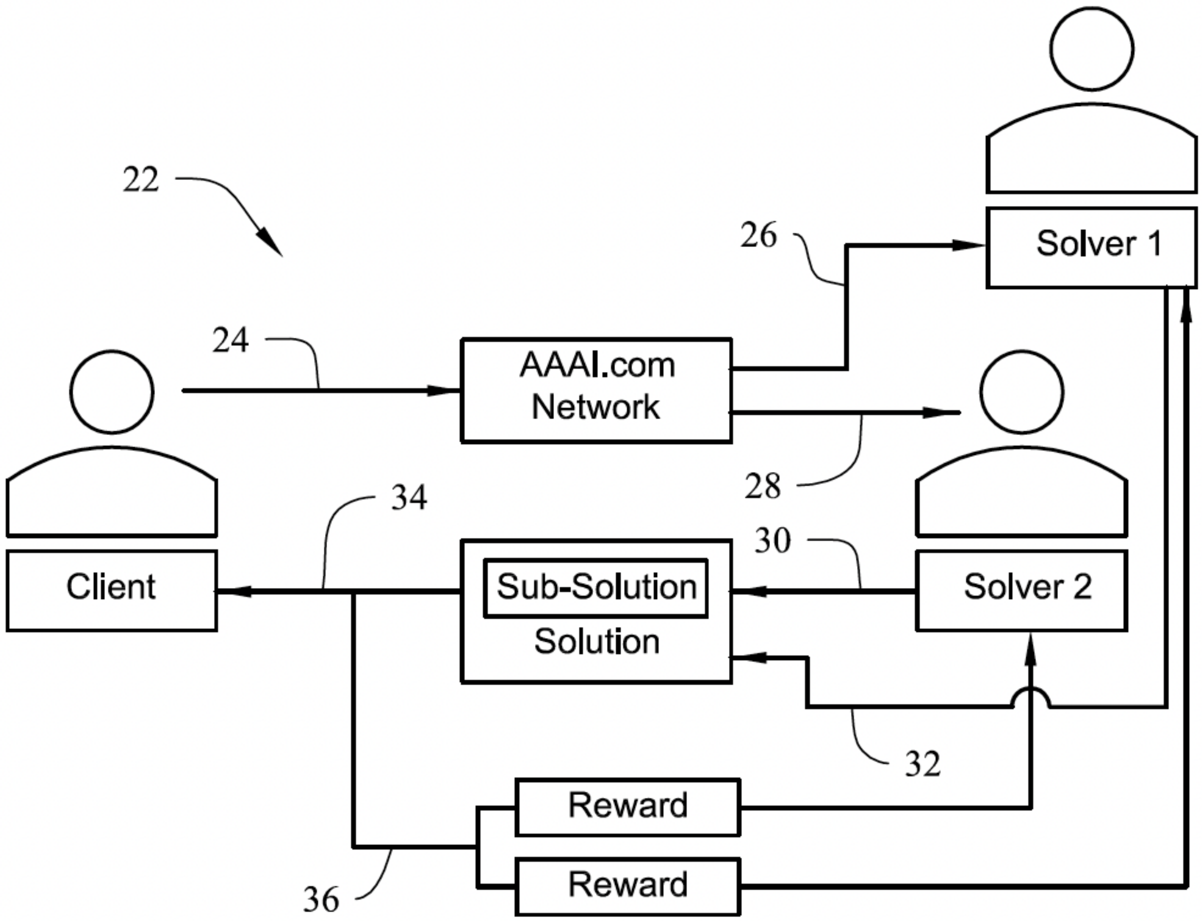

FIG. 12

FIGURE 13: AAAIS AND CLONES PARTICIPATING IN MARKETPLACE

A schematic block diagram of an exemplary utilization of multiple customized AAAIs and their cloned AAAIs participating in an AAAI marketplace over a network.

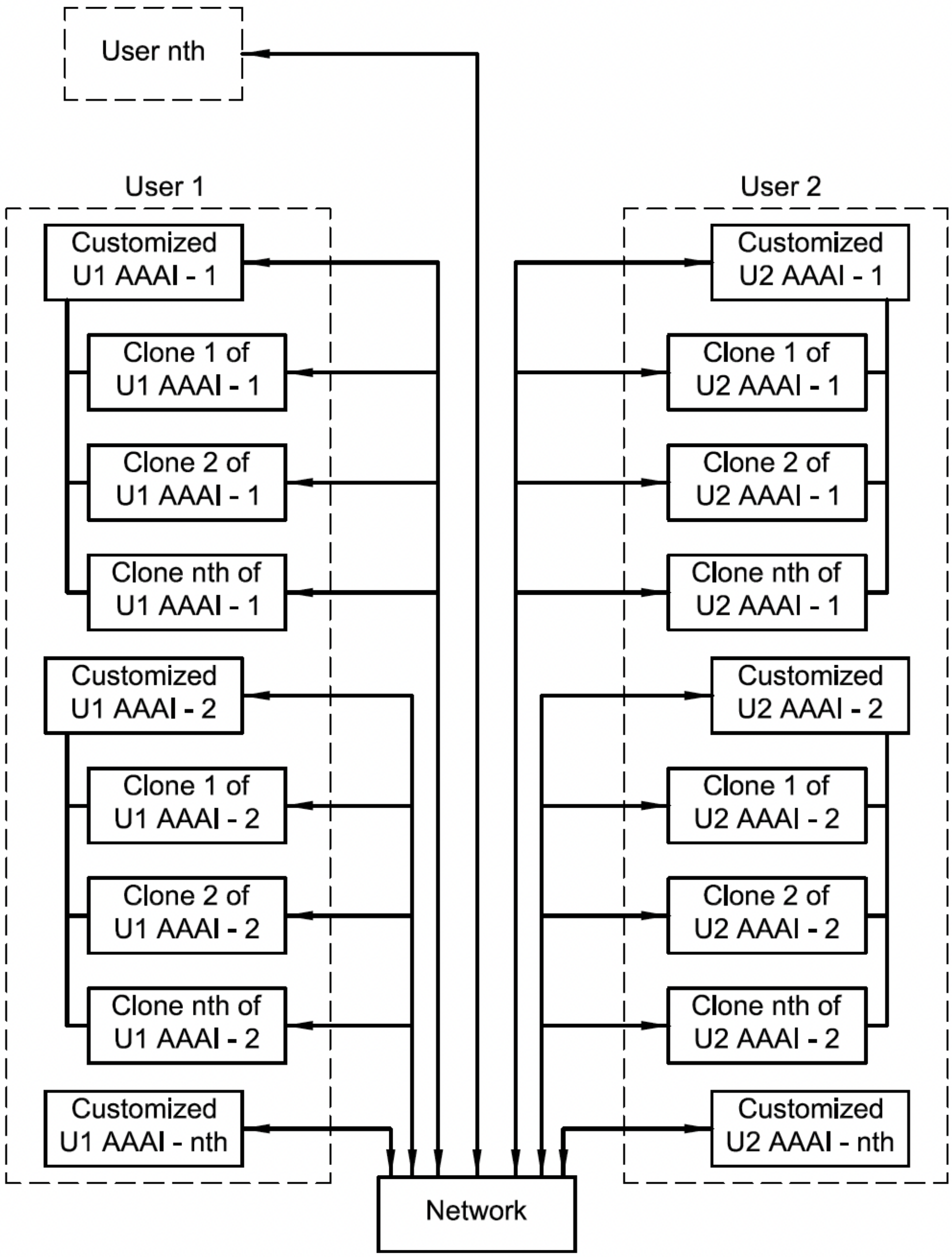

FIG. 13

FIGURE 14: METHODS FOR IMPLEMENTING PROBLEM-SOLVING SYSTEM

A diagram illustrating features and functions of the Problem-Solving architecture, including the Tree structure used by the WorldThink protocol.

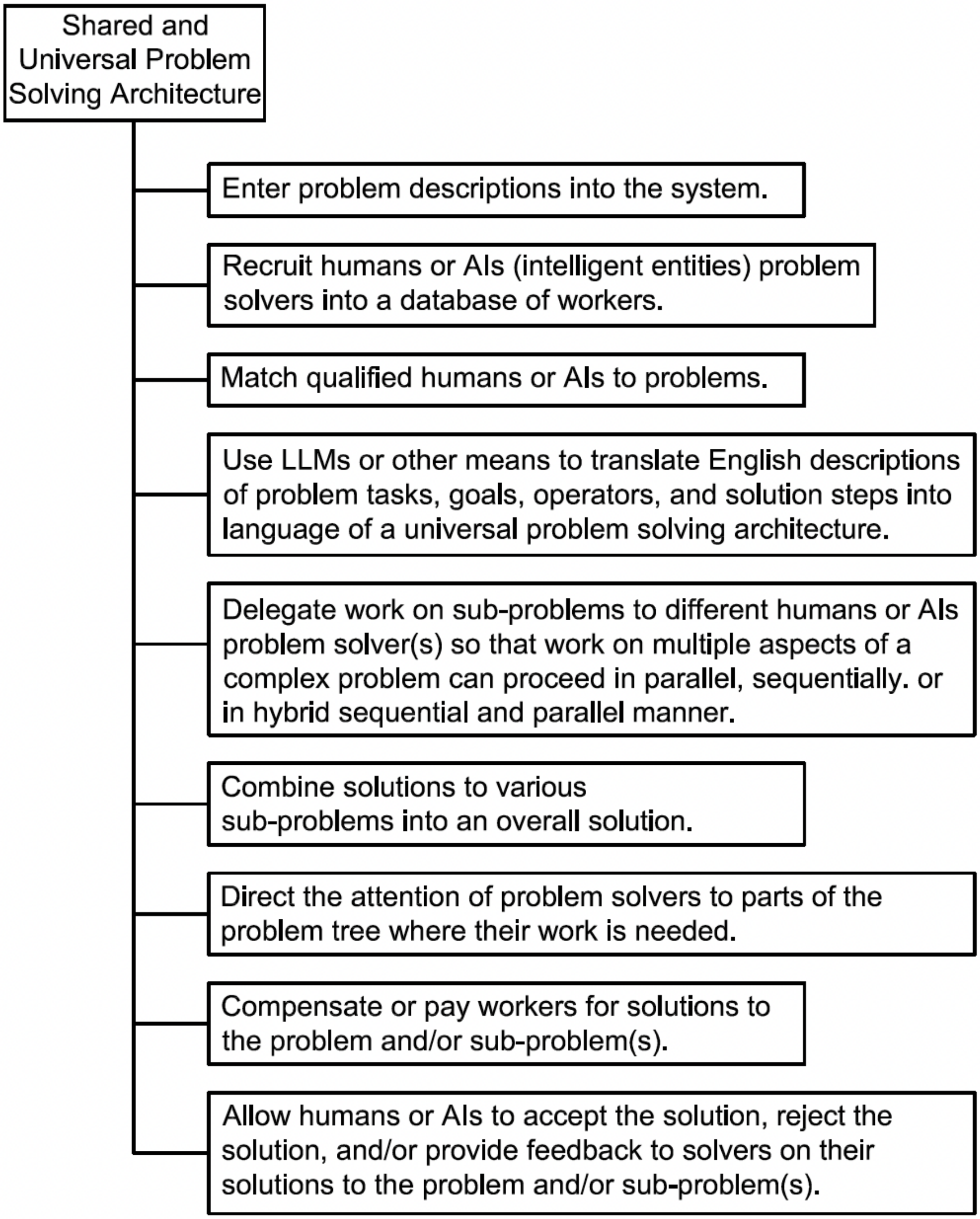

FIG. 14

FIGURE 15: PROBLEM-SOLVING USING COMMON COGNITIVE ARCHITECTURE

A flow chart illustrating an exemplary problem-solving process utilizing a typical cognitive architecture implemented in an AI system.

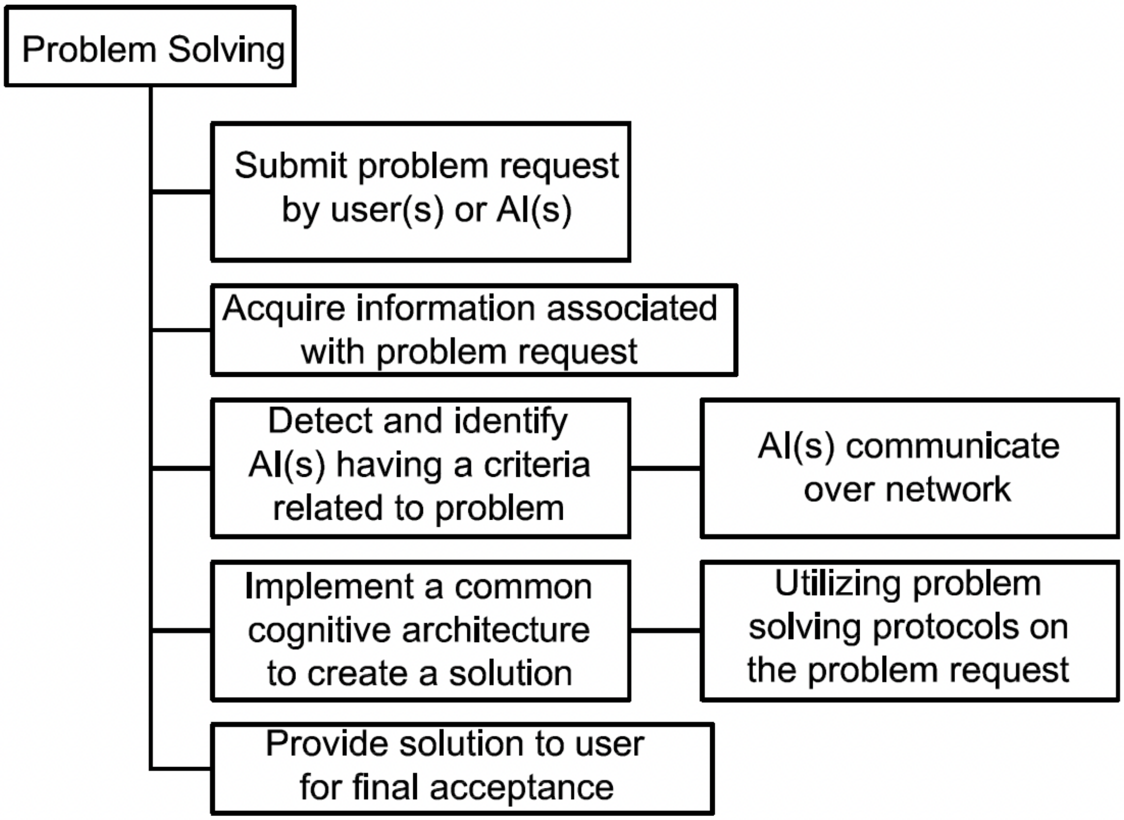

FIG. 15

FIGURE 16: AAAI PROBLEM-SOLVING PROCESS

A flow chart illustrating an exemplary embodiment of the AAAI problem-solving process of the present technology.

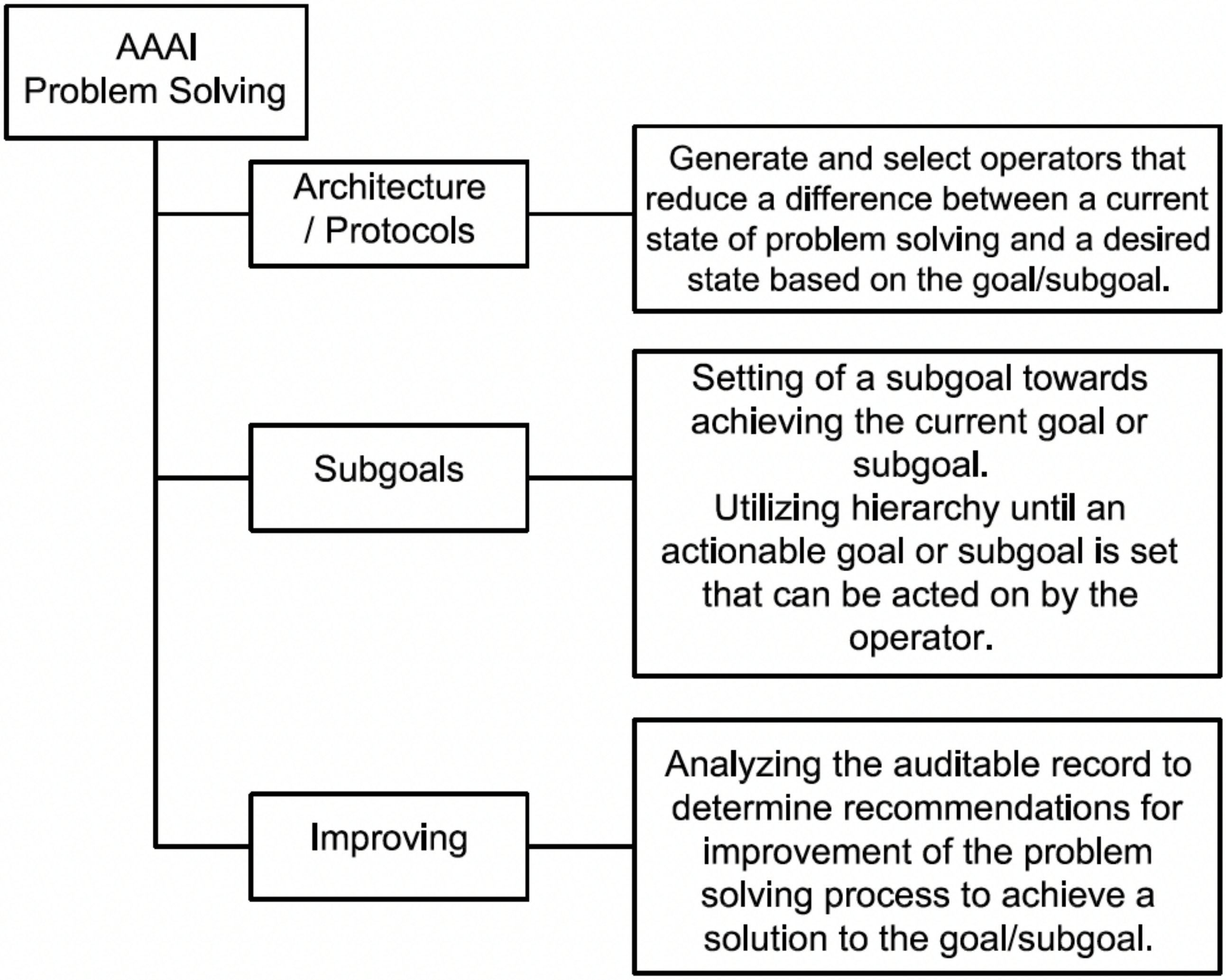

FIG. 16

FIGURE 17: FEATURES AND FUNCTIONS OF PROBLEM-SOLVING TREE

A diagram illustrating features and functions of the Problem-Solving Tree structure used in the WorldThink protocol.

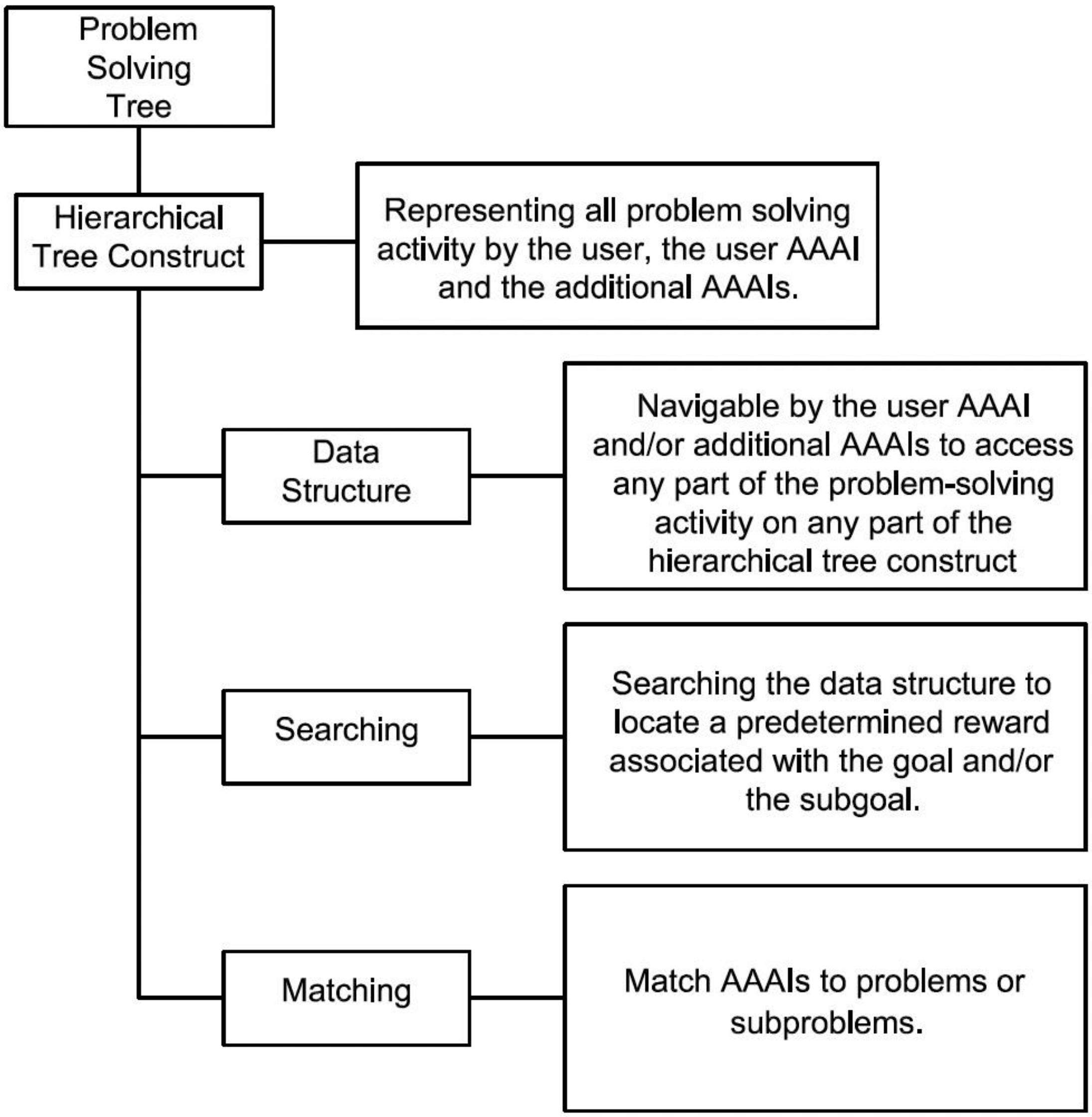

FIG. 17

FIGURE 18: SAFETY AND ETHICS CHECK PROCESS

A flow chart illustrates an exemplary embodiment of the present technology's safety/ethics check process.

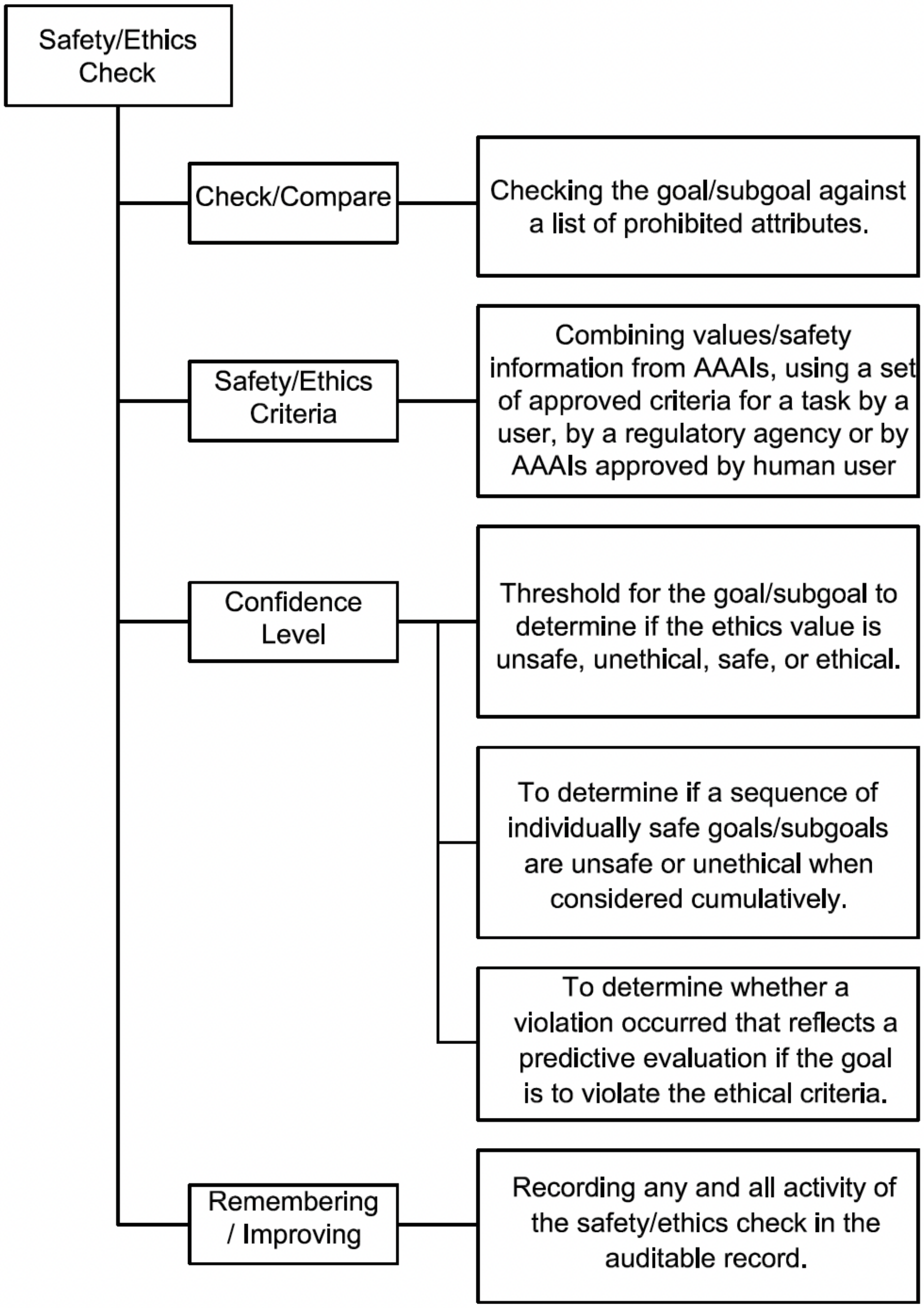

FIG. 18

FIGURE 19: SAFETY AND ETHICS CHECKS WITH TRIGGERS

A flow chart illustrating an exemplary embodiment of the safety and ethics checks subsystem or process, including triggering mechanisms.

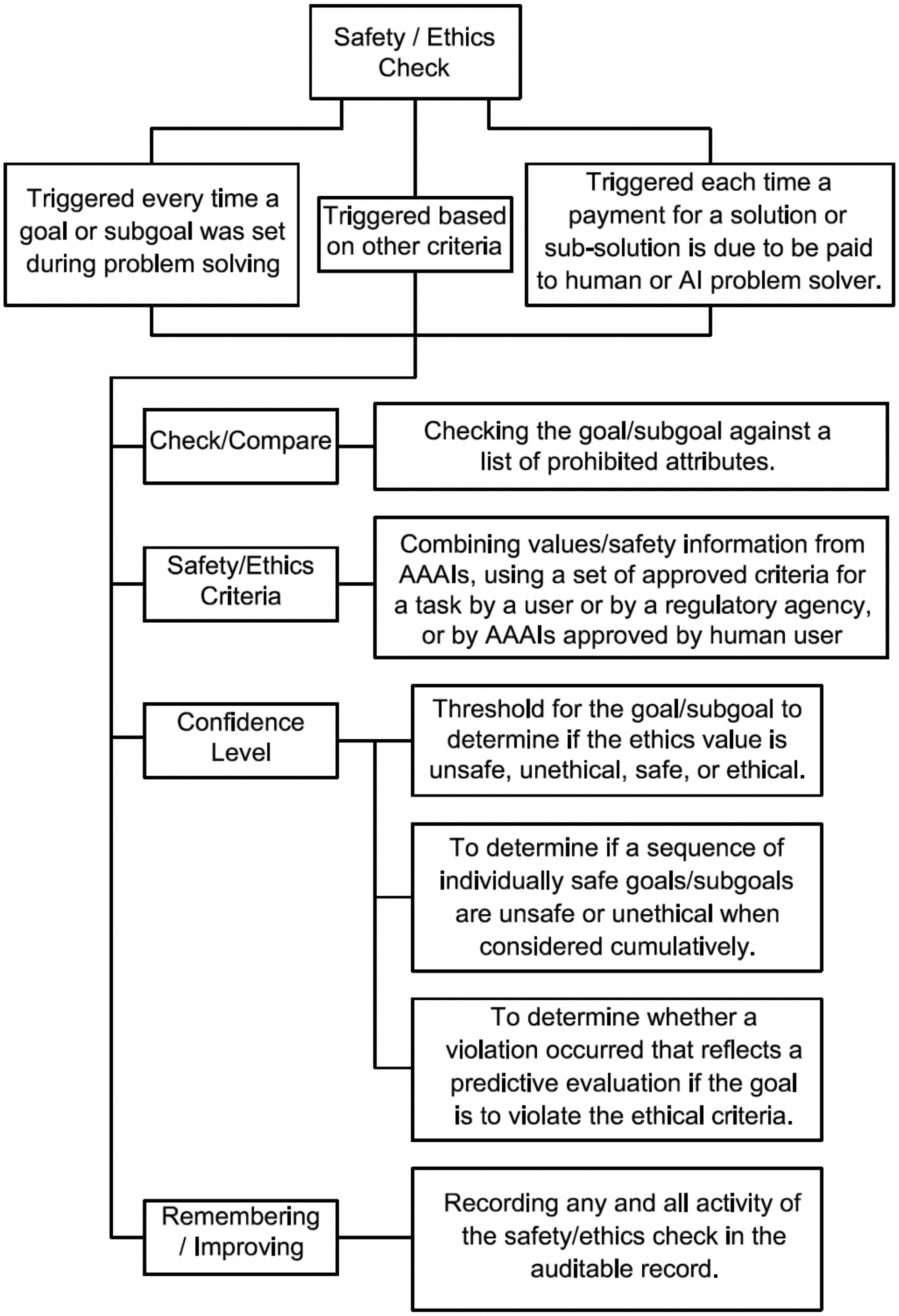

FIG. 19

FIGURE 20: RECORDING AND IMPROVING PROCESS

A flow chart illustrating an exemplary embodiment of the recording/improving process of the present technology.

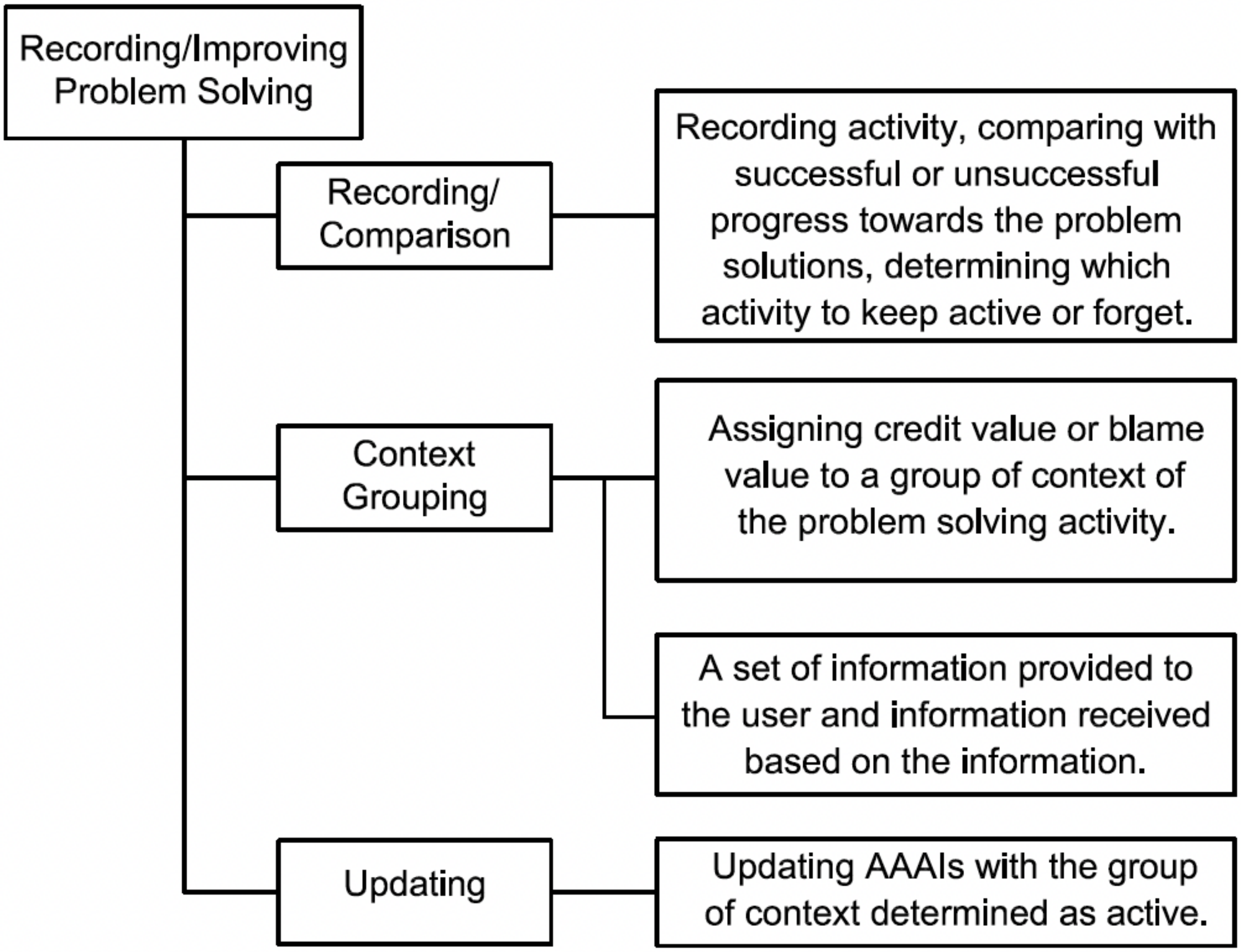

FIG. 20

FIGURE 21: NATURAL LANGUAGE TO PROBLEM-SOLVING TRANSLATION

A flow chart illustrating an exemplary embodiment of the natural language to problem-solving language translator subsystem or process.

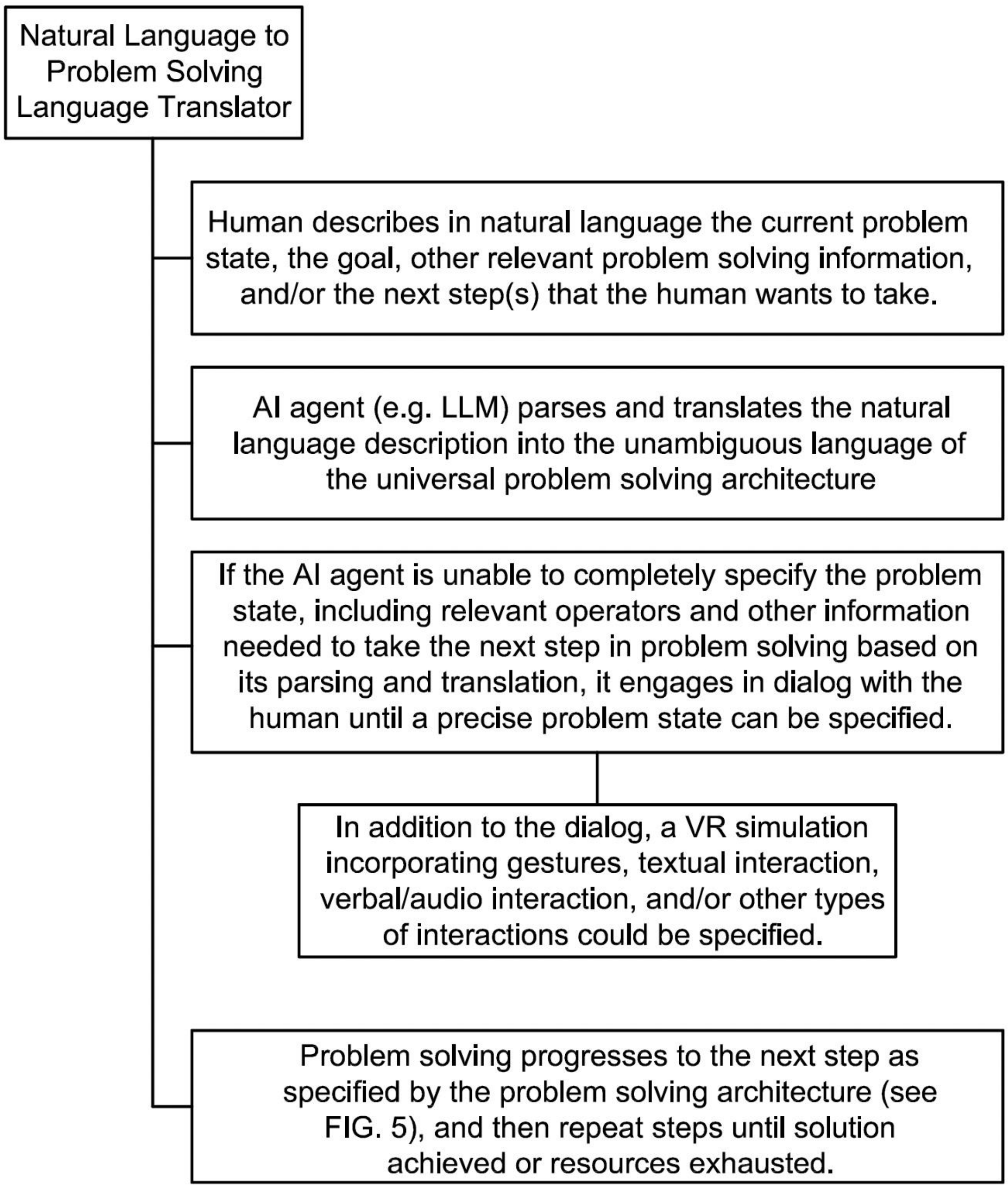

FIG. 21

FIGURE 22: CUSTOMIZATION PROCESS FOR AAAI OR AI

A flow chart illustrating an exemplary customization process of an AAAI or other advanced AI.

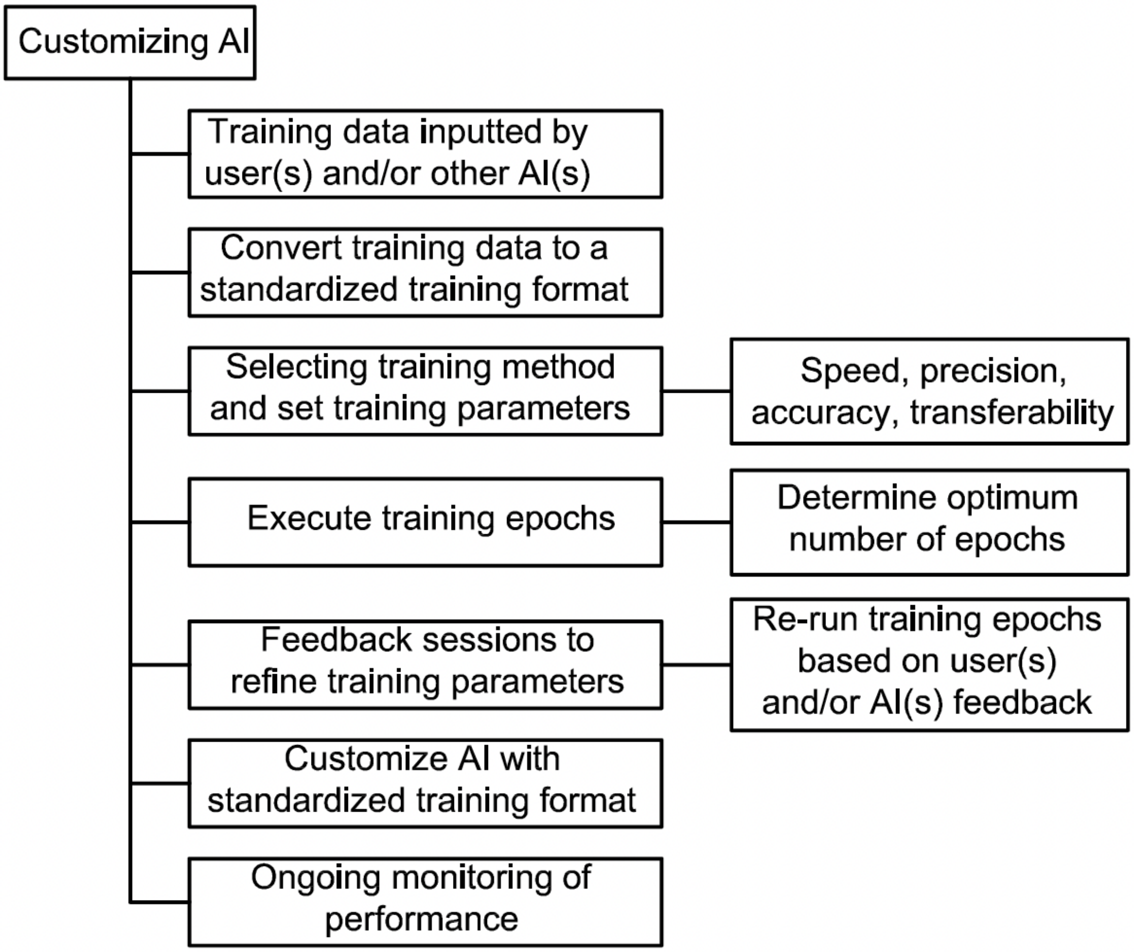

FIG. 22

FIGURE 23: ELICITING ETHICAL INFORMATION FROM ENTITIES

A flow chart illustrating an exemplary process including detailed methods of eliciting ethical information or preferences for the intelligent entities.

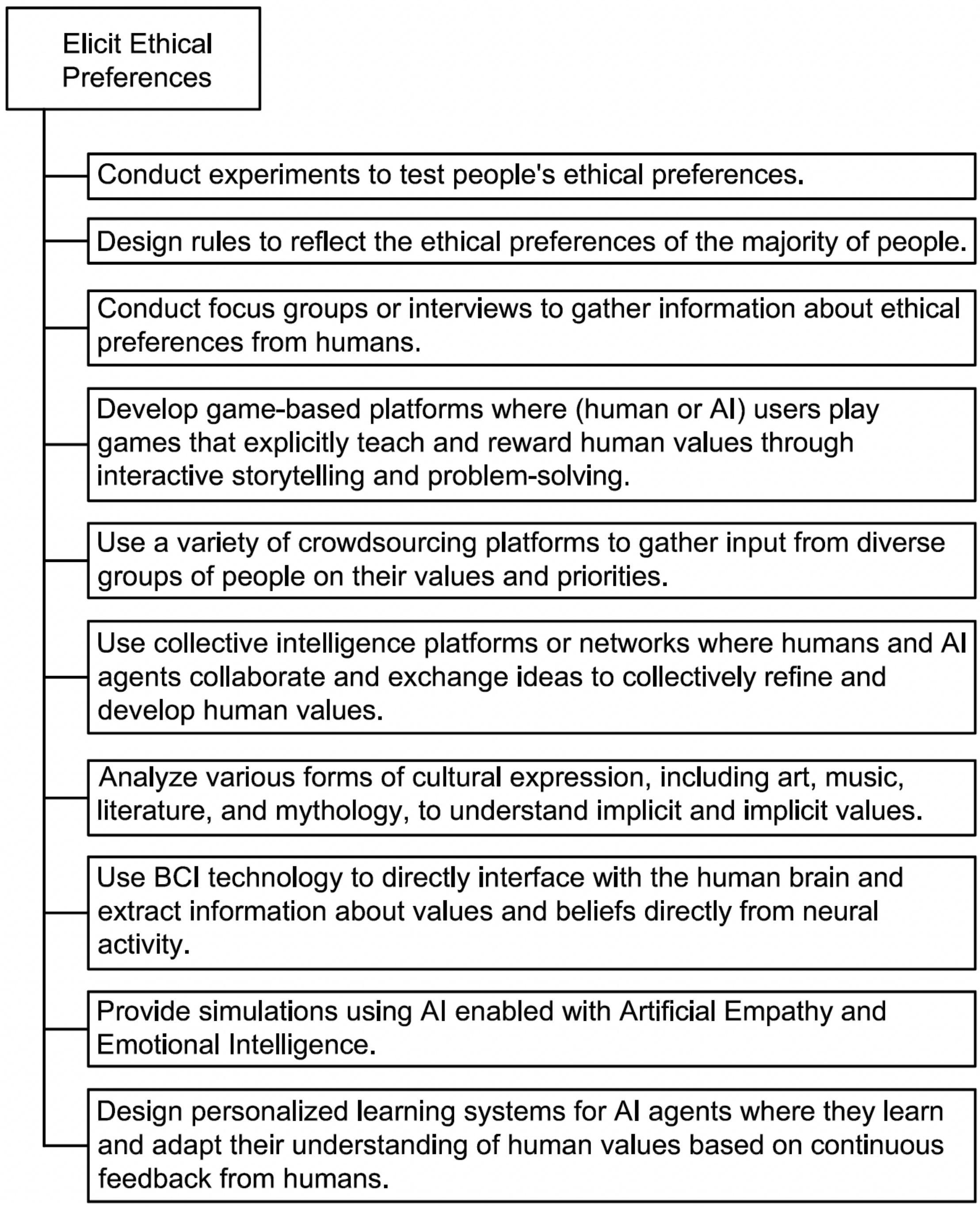

FIG. 23

FIGURE 24: AUTOMATED QUESTIONNAIRES FOR ETHICAL INPUT

A flow chart illustrating an exemplary automatic generation process and use of questionnaires provided to the intelligent entities.

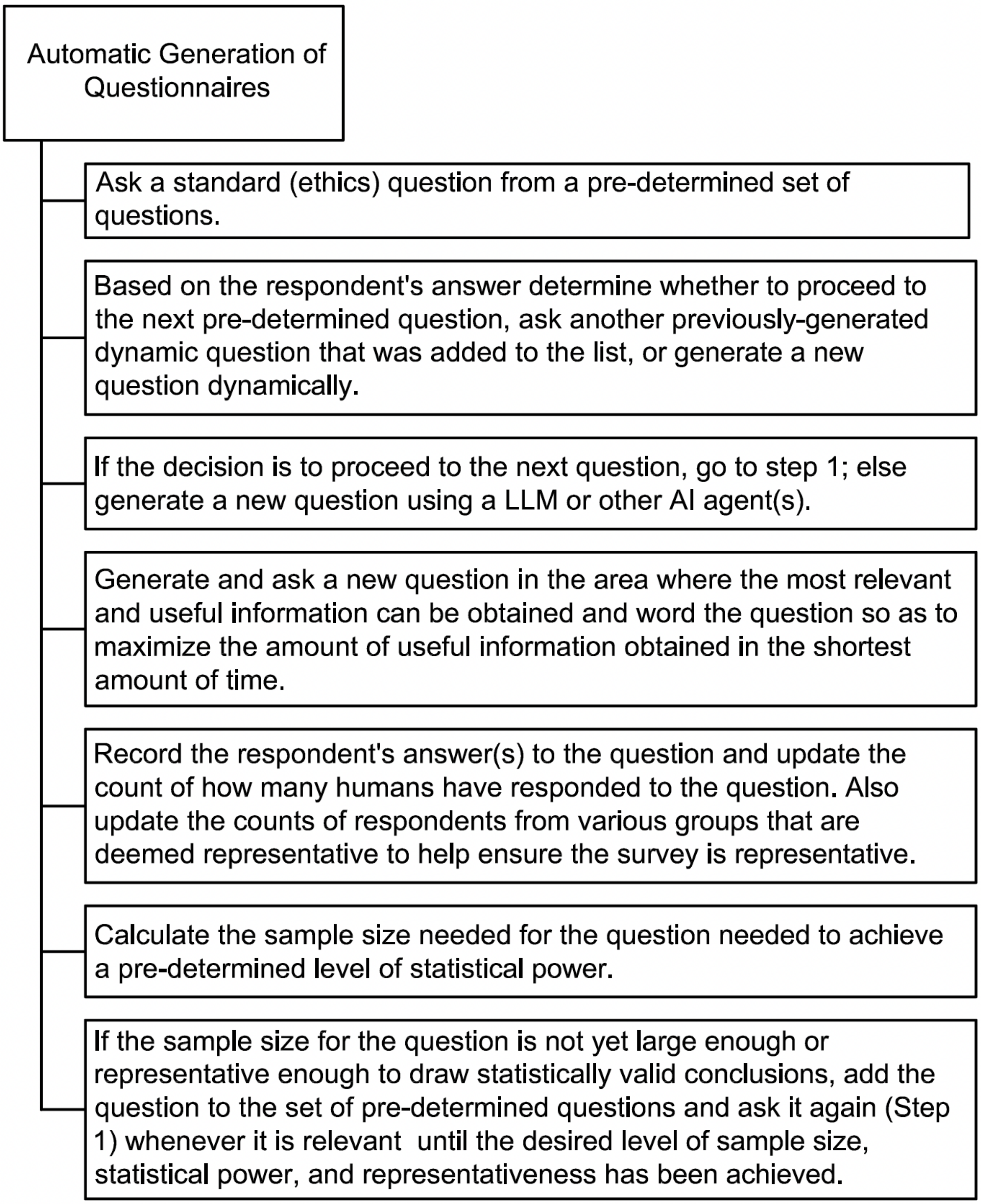

FIG. 24

FIGURE 25: INDUCING ETHICAL VALUES FROM HUMAN BEHAVIOR
A flow chart illustrating an exemplary process of inducing ethical values by detecting patterns in human behavior.

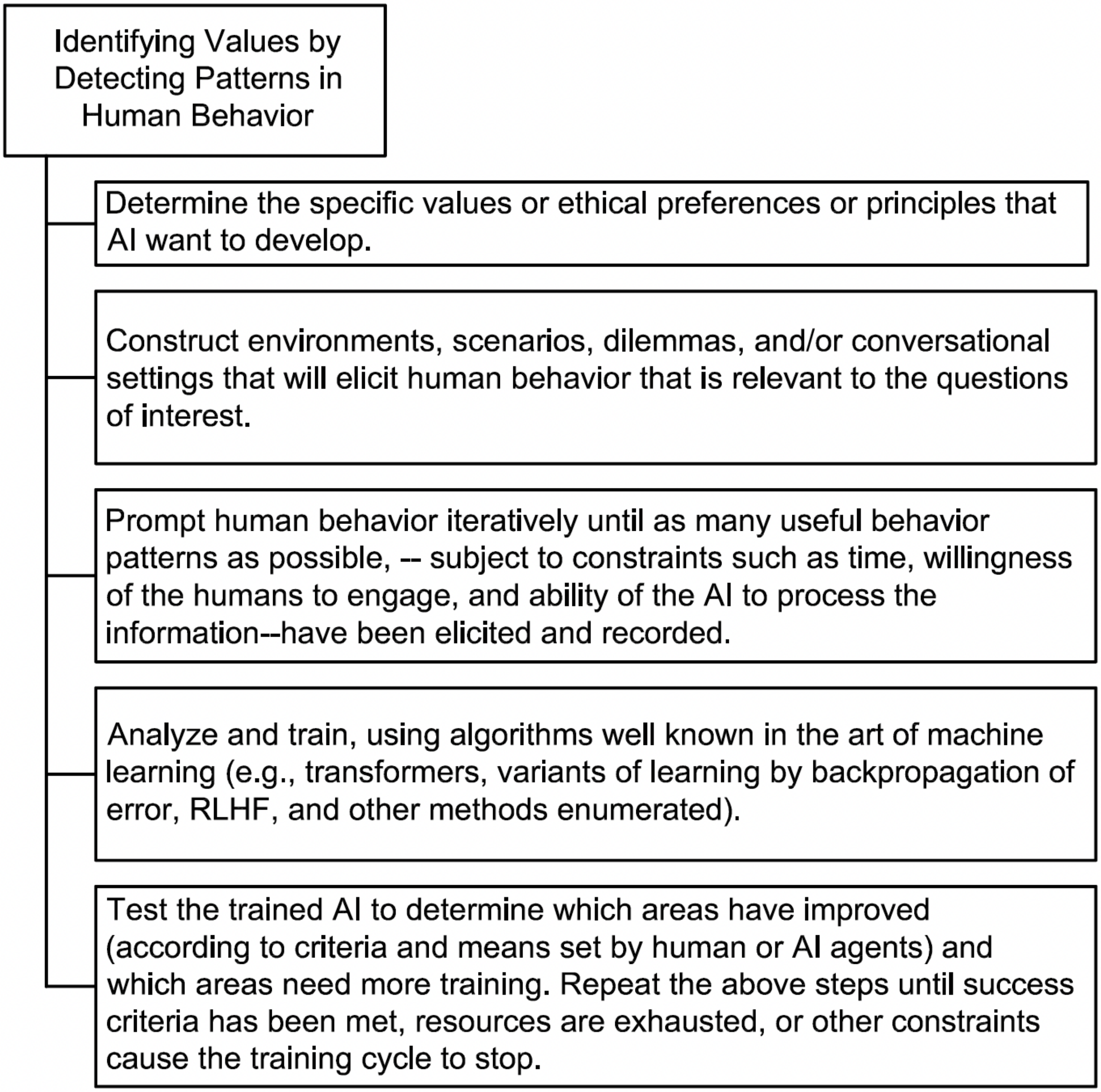

FIG. 25

FIGURE 26: CUSTOMIZING AI TO INDIVIDUAL ETHICAL PROFILES

A flow chart illustrating an exemplary embodiment of the process for customizing AI, such as Base LLMs, so that the AI incorporates ethical and other information from specific human users or to each user's ethical or informational profile.

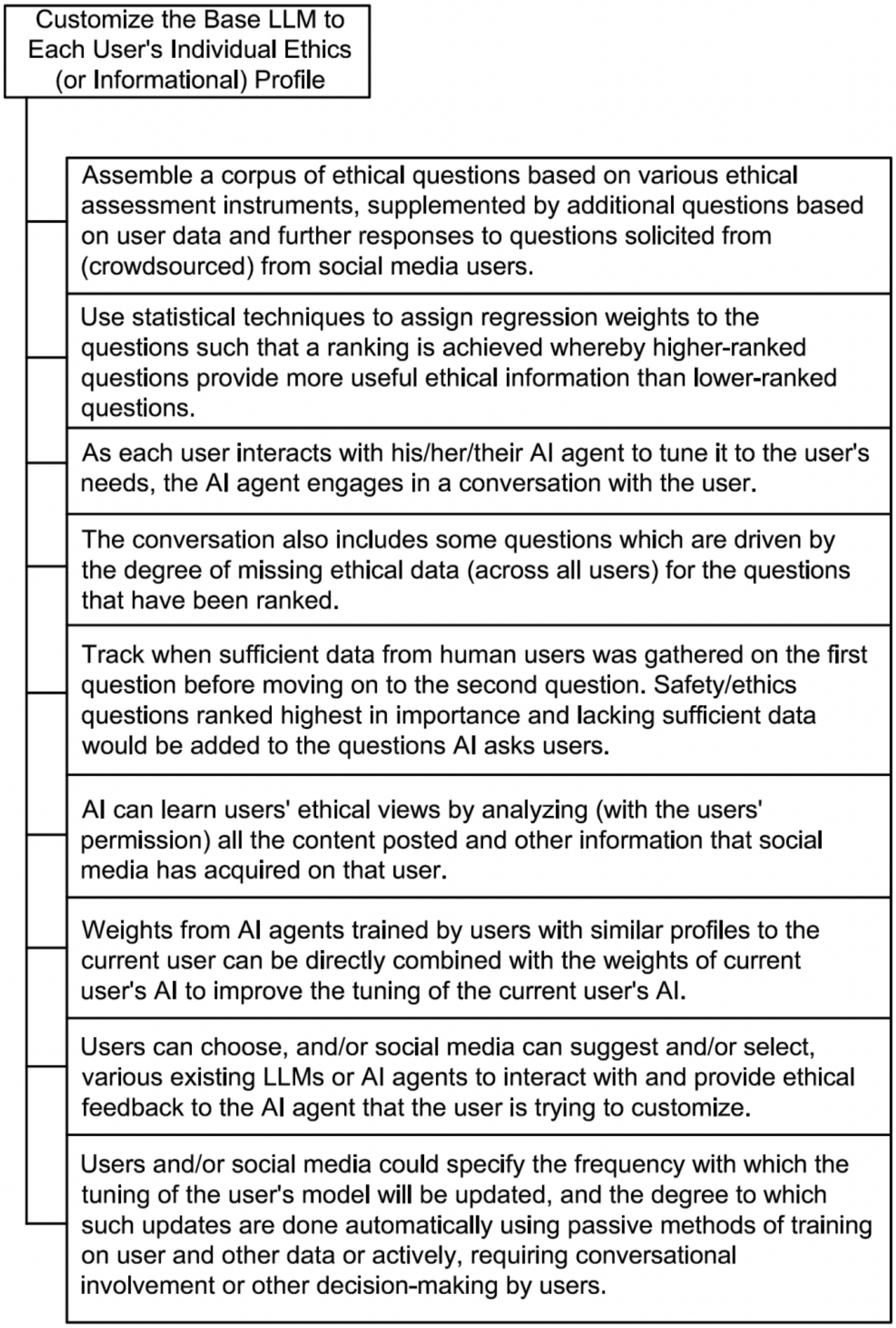

FIG. 26

FIGURE 27: TRAINING AI WITH SAFETY AND ETHICAL GUARDRAIL

A flow chart illustrating an exemplary embodiment of the process for training an AI, such as a Base LLM, to incorporate safety and ethical guardrails for safe and scalable AGI.

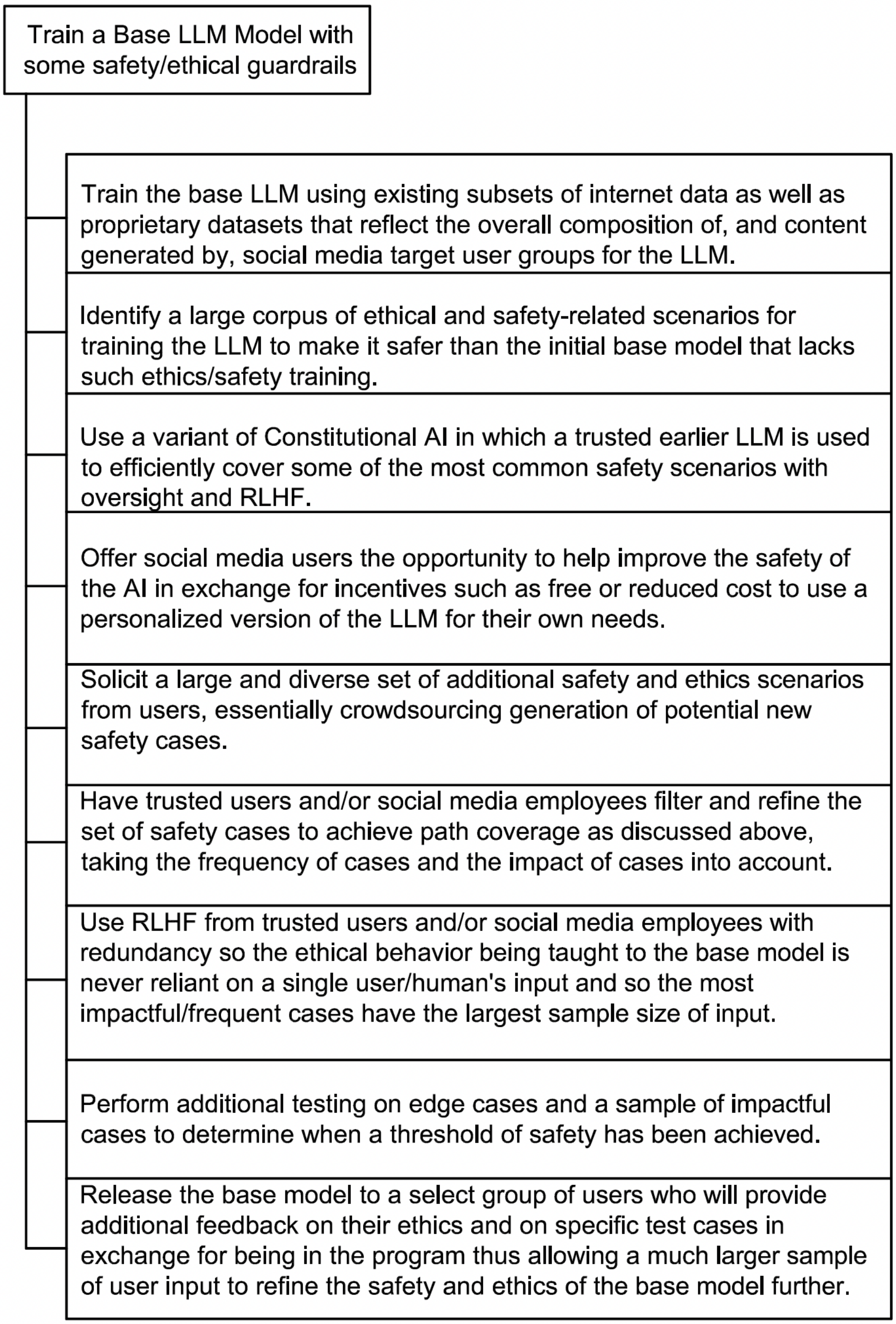

FIG. 27

FIGURE 28: CREATING SCALABLE, ETHICAL AGI WITH CUSTOMIZED AIS

A flow chart illustrating an exemplary embodiment of the general overall process of the present technology for creating scalable, ethical AGI using the customized AIs, each possessing ethical information.

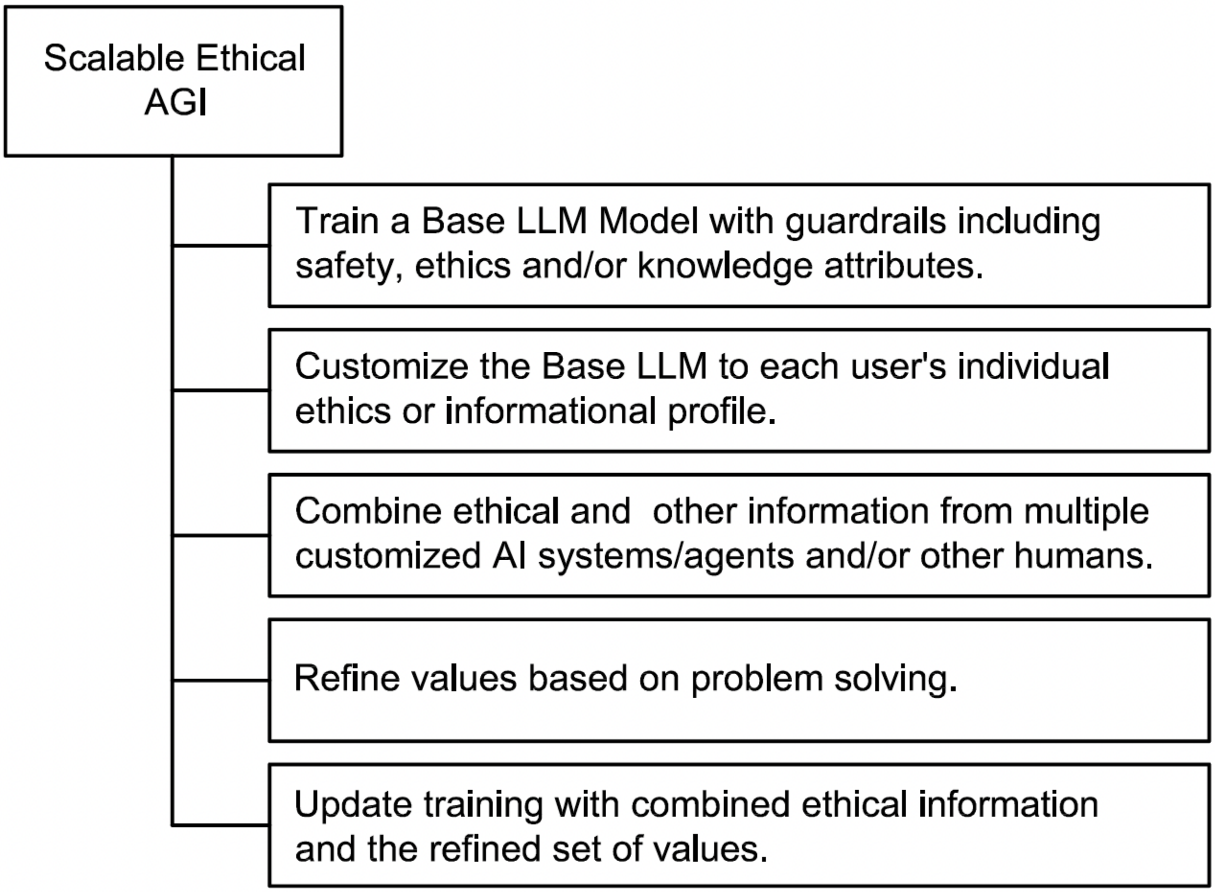

FIG. 28

FIGURE 29: REPUTATIONAL COMPONENT FOR SAFETY AND ETHICS

A flow chart illustrating an exemplary embodiment of the reputational component subsystem or process for AI systems or agents that can be used to enhance the effectiveness, safety, and ethics of the system.

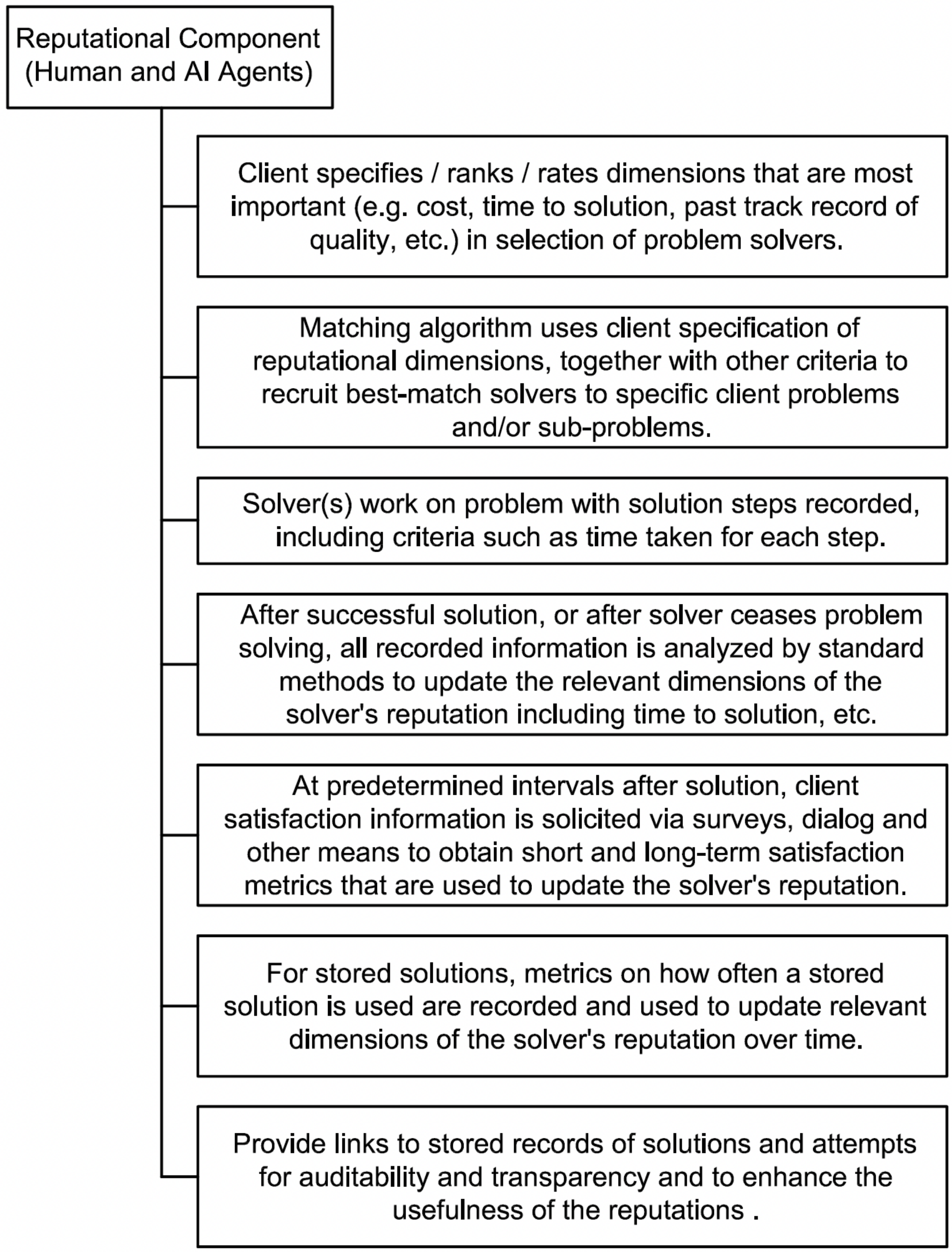

FIG. 29

FIGURE 30: CUSTOMIZATION AND CROSS-PLATFORM PROCESS

A flow chart illustrates an exemplary embodiment of the present technology's customization process and the cross-platform process.

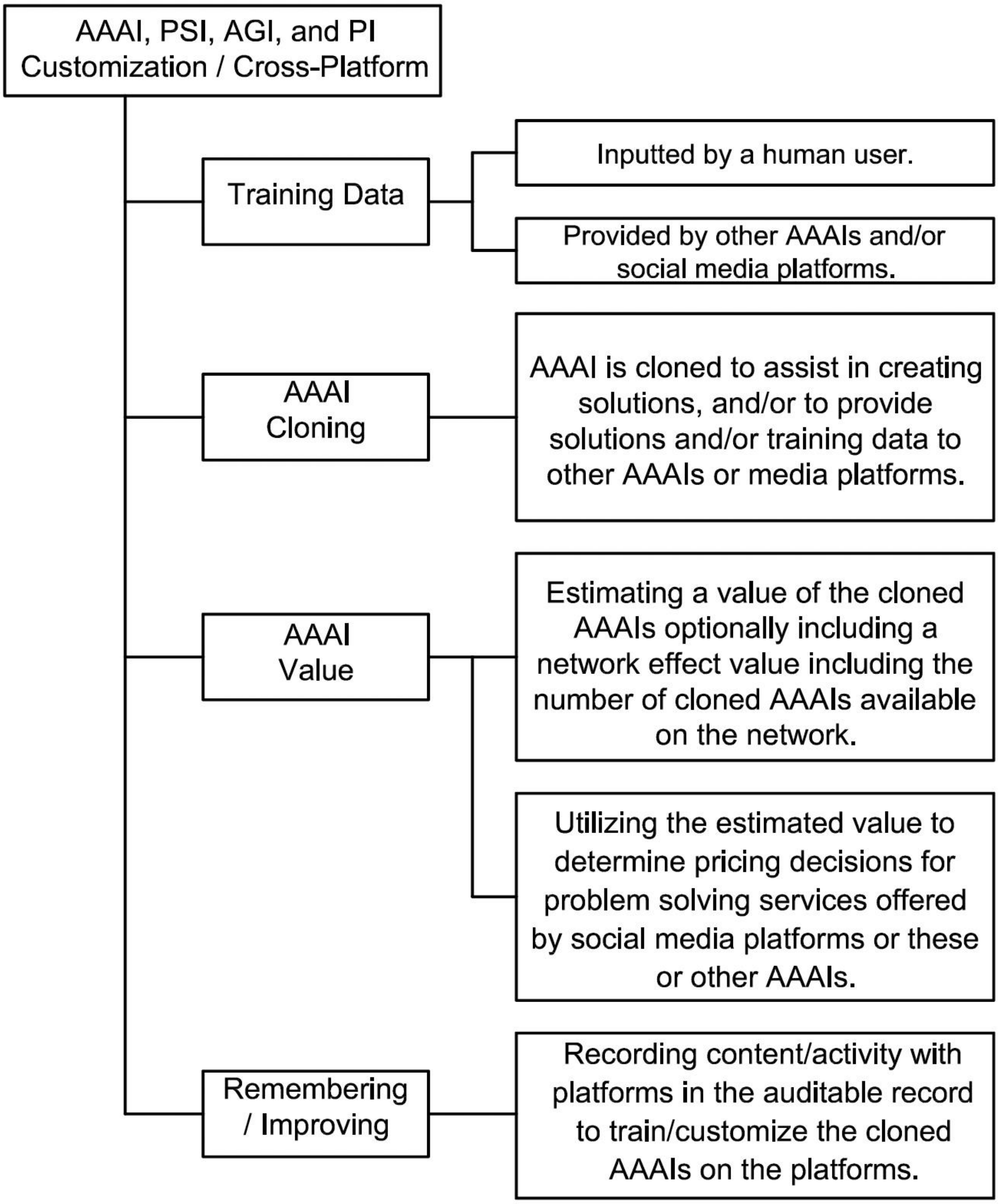

FIG. 30

FIGURE 31: ADDITIONAL CUSTOMIZATION OF AI FOR HUMAN-CENTERED DESIGN

A flow chart illustrating an exemplary embodiment of additional customization.

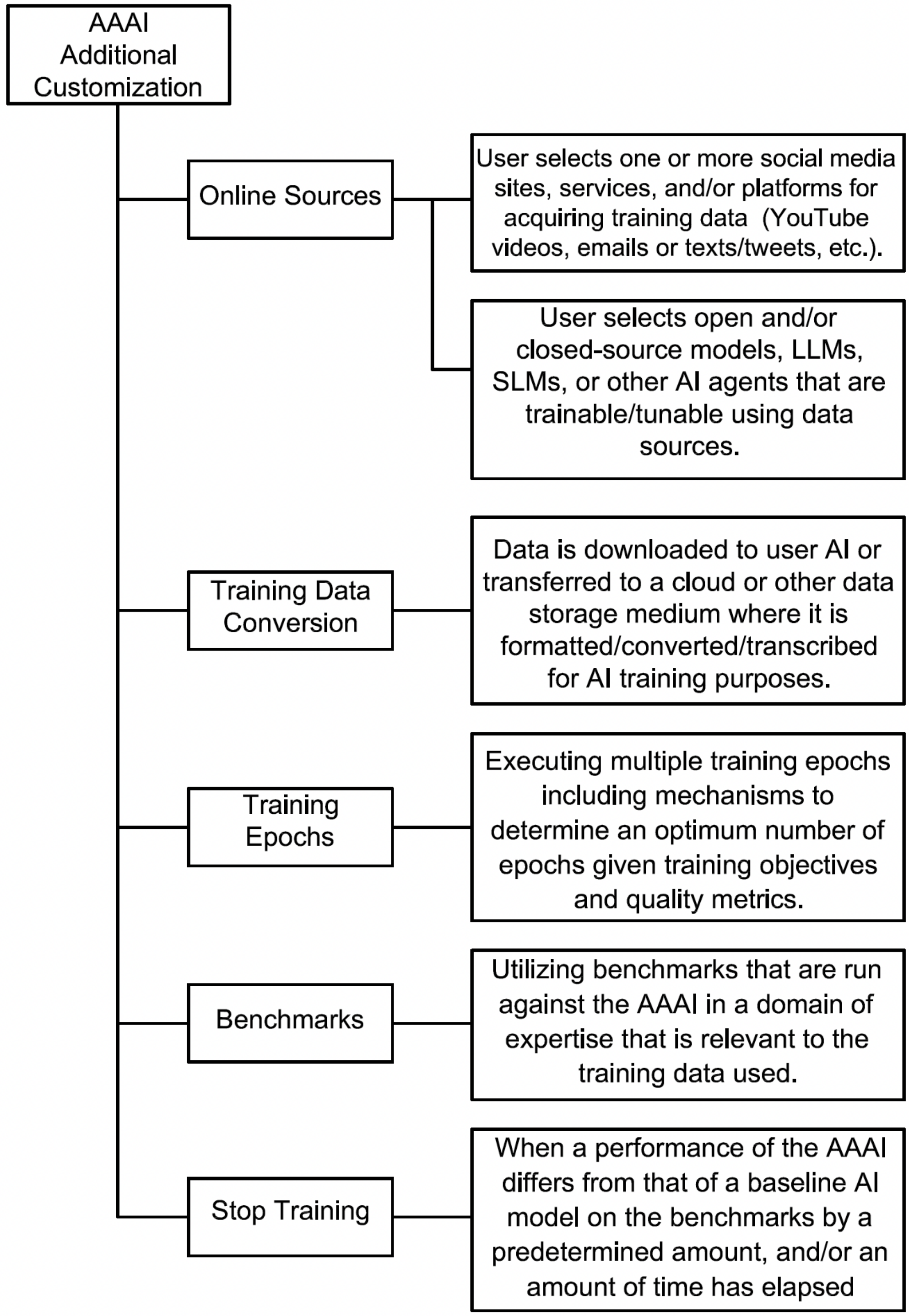

FIG. 31

FIGURE 32: COMBINING ETHICAL INFORMATION FROM MULTIPLE AGENTS

A flow chart illustrating an exemplary embodiment of the process for combining ethical or other information from multiple customized AI agents.

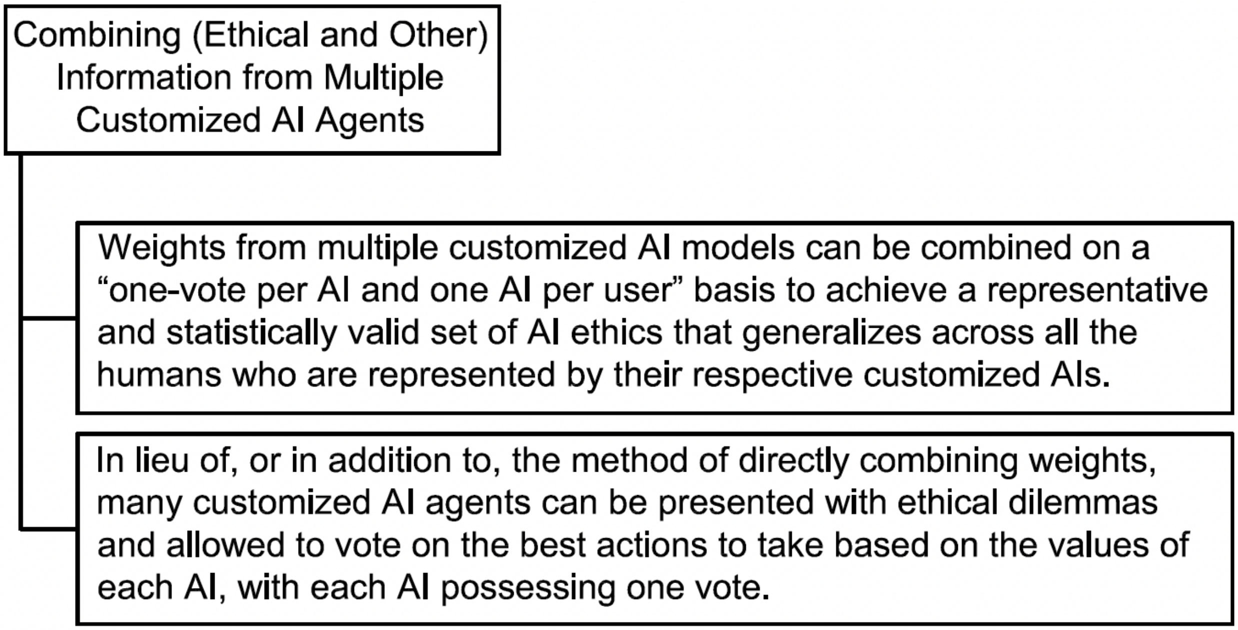

FIG. 32

FIGURE 33: REFINING VALUES BASED ON PROBLEM-SOLVING

A flow chart illustrating an exemplary embodiment of the process for refining values utilized in customizing and based on problem-solving.

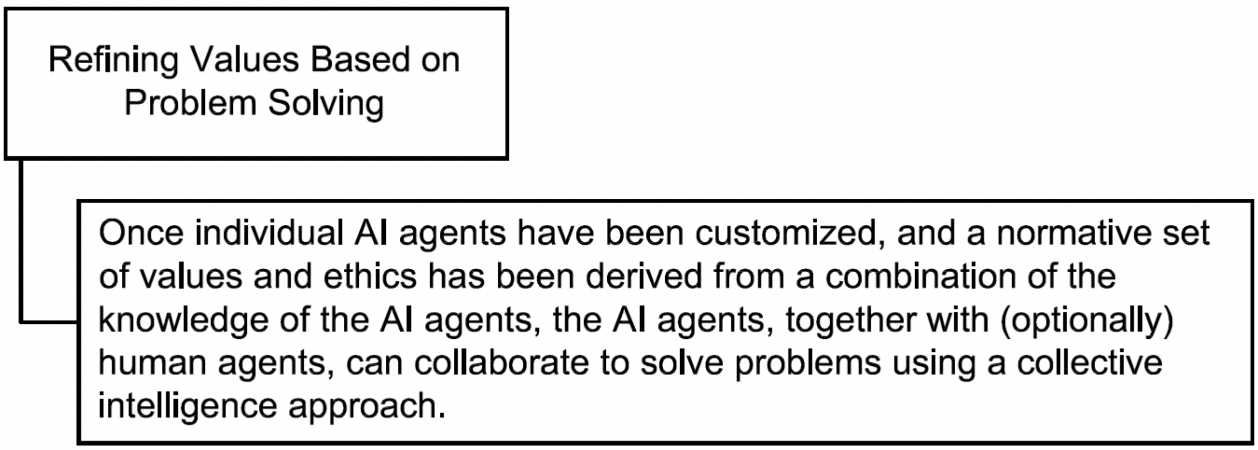

FIG. 33

FIGURE 34: SAFE, SCALABLE AGI USING WEIGHT MATRICES

A flow chart illustrating an exemplary embodiment of the process for creating safe and scalable AGI, applicable also to scalable PI, using combinations of weight matrices from multiple identified AI agents.

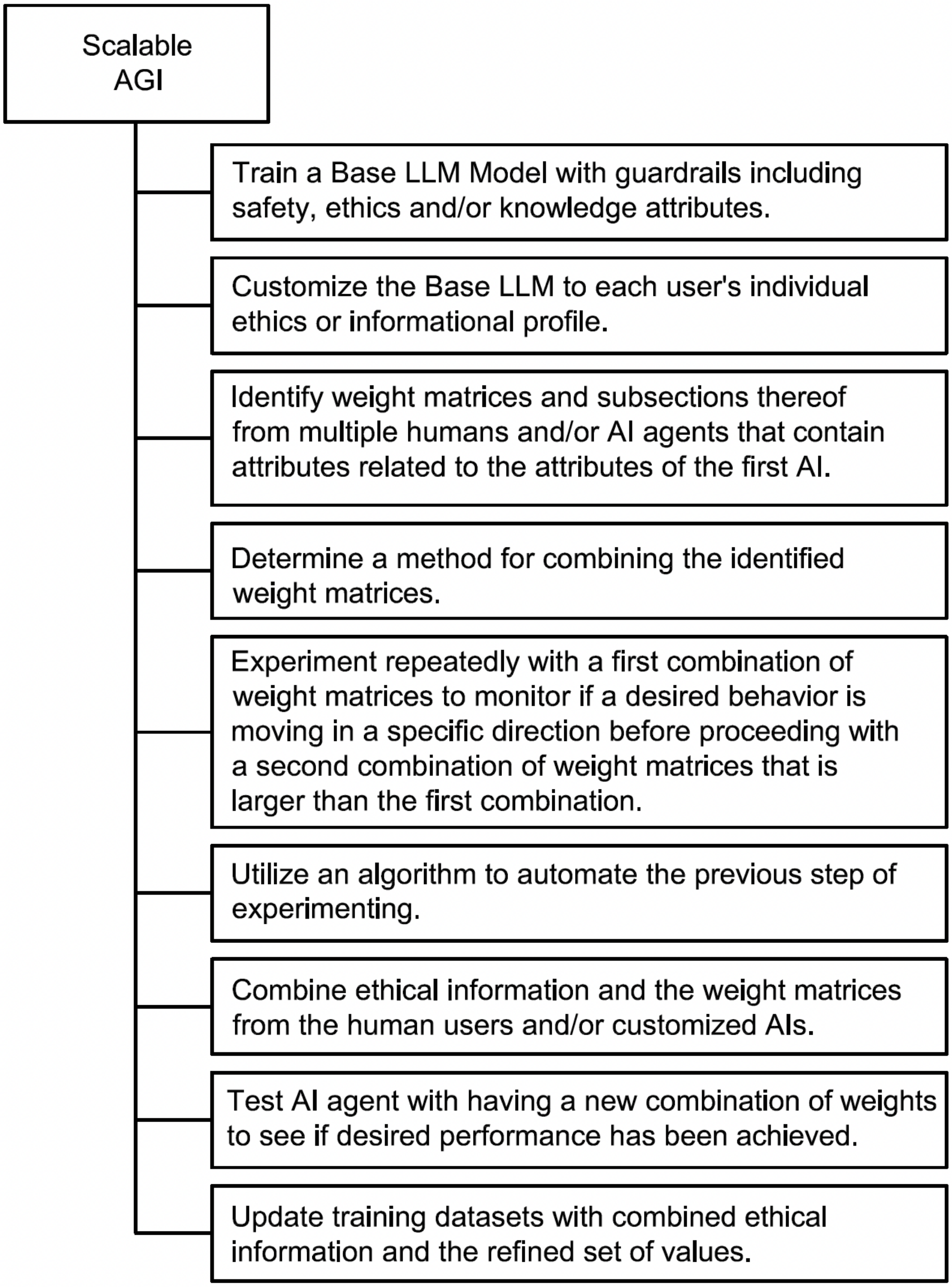

FIG. 34

FIGURE 35: SAFE, SCALABLE AGI WITH MONITORING ETHICAL ISSUES

A flow chart illustrating an exemplary embodiment of the process for creating safe and scalable AGI, applicable also to scalable PI, using combinations of weight matrices from multiple identified AI agents in combination with monitoring and flagging potential ethical issues.

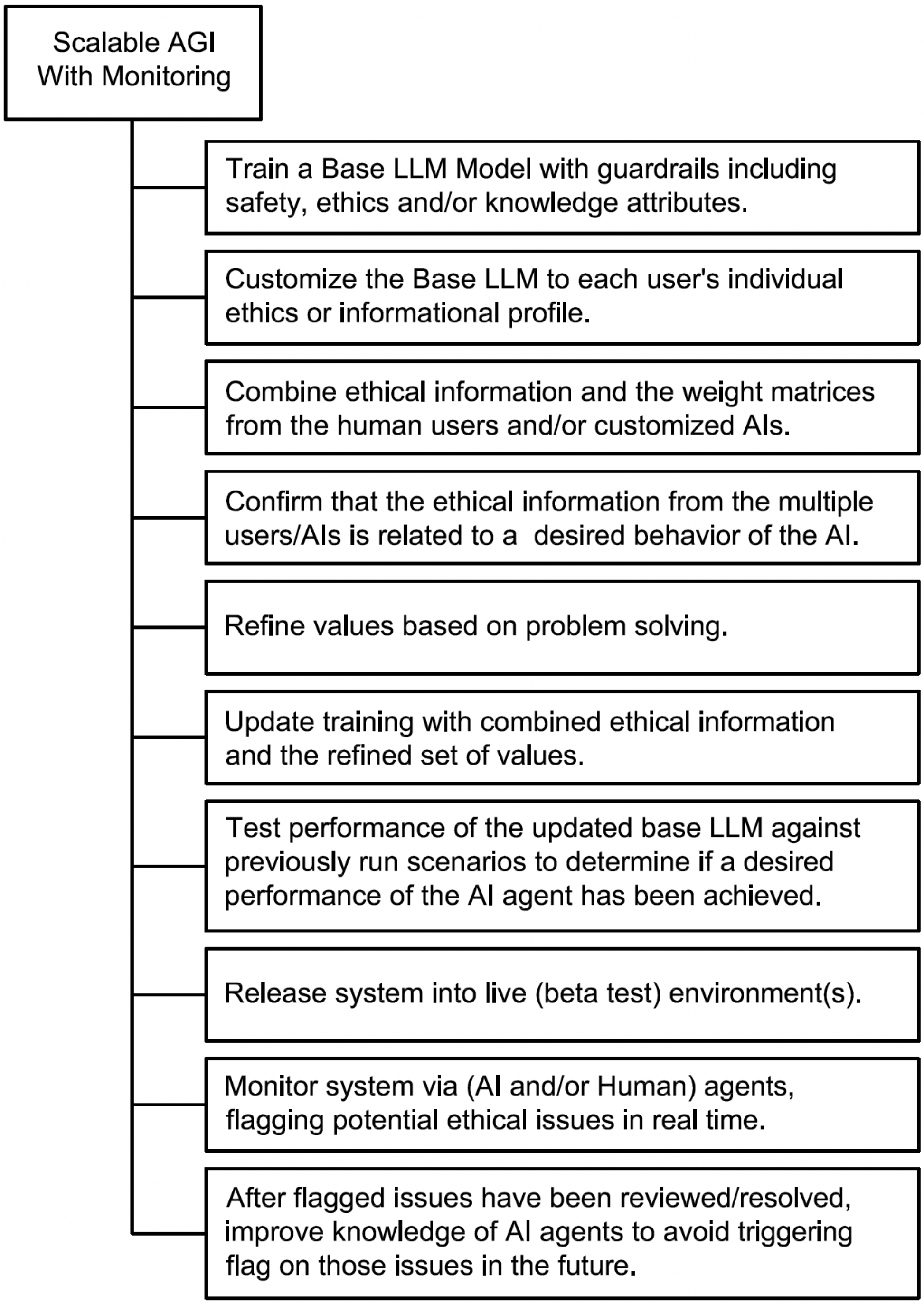

FIG. 35

FIGURE 36: PREVENTING LLM HALLUCINATION

A flow chart illustrating an exemplary embodiment of the process for preventing hallucination by LLMs in the present technology.

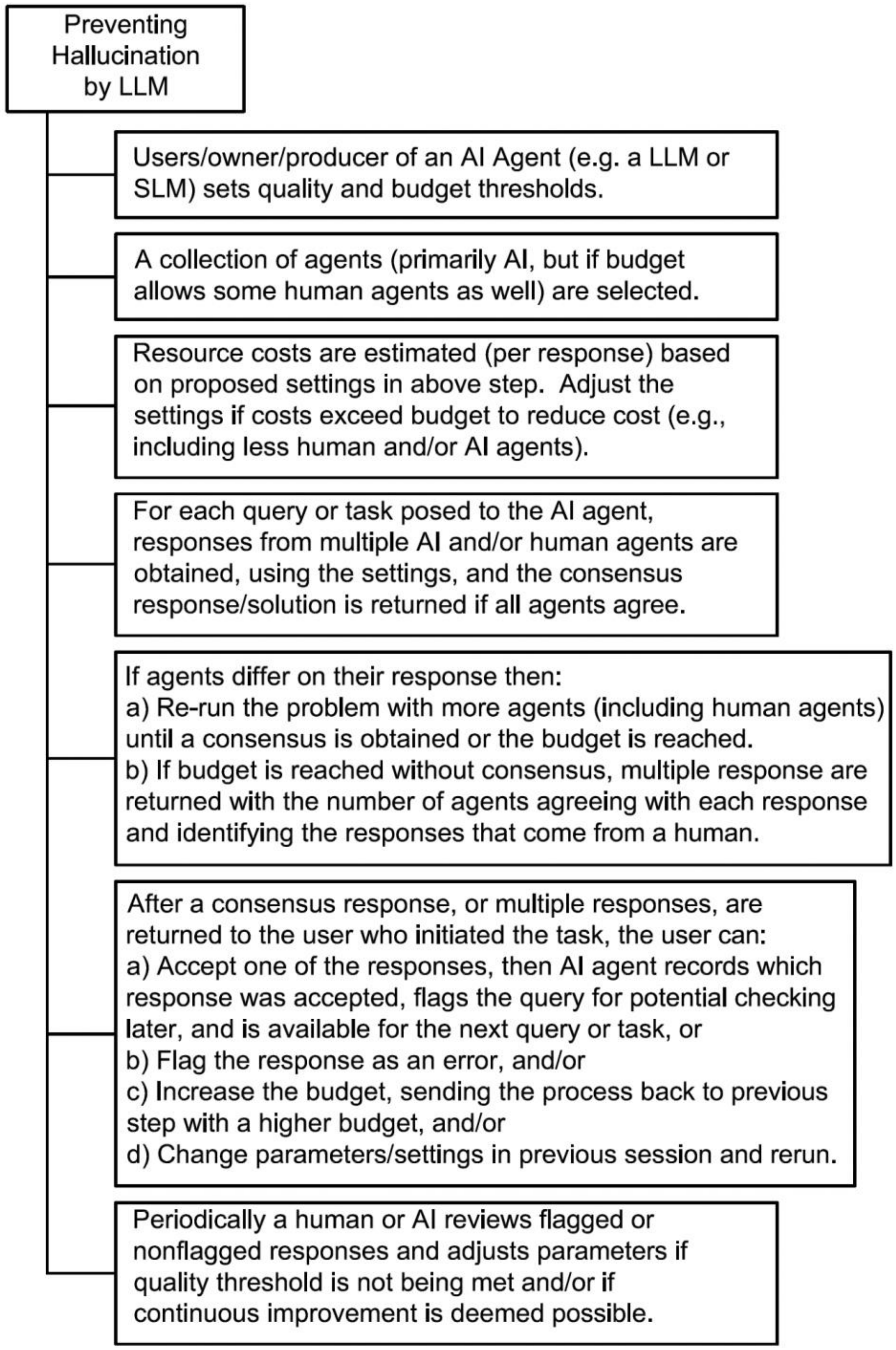

FIG. 36

FIGURE 37: KNOWLEDGE MODULES TO CUSTOMIZE SYSTEMS

A flow chart illustrating an exemplary embodiment of the process for the use of knowledge modules or collections of agents to customize the AI, AGI, or PI systems.

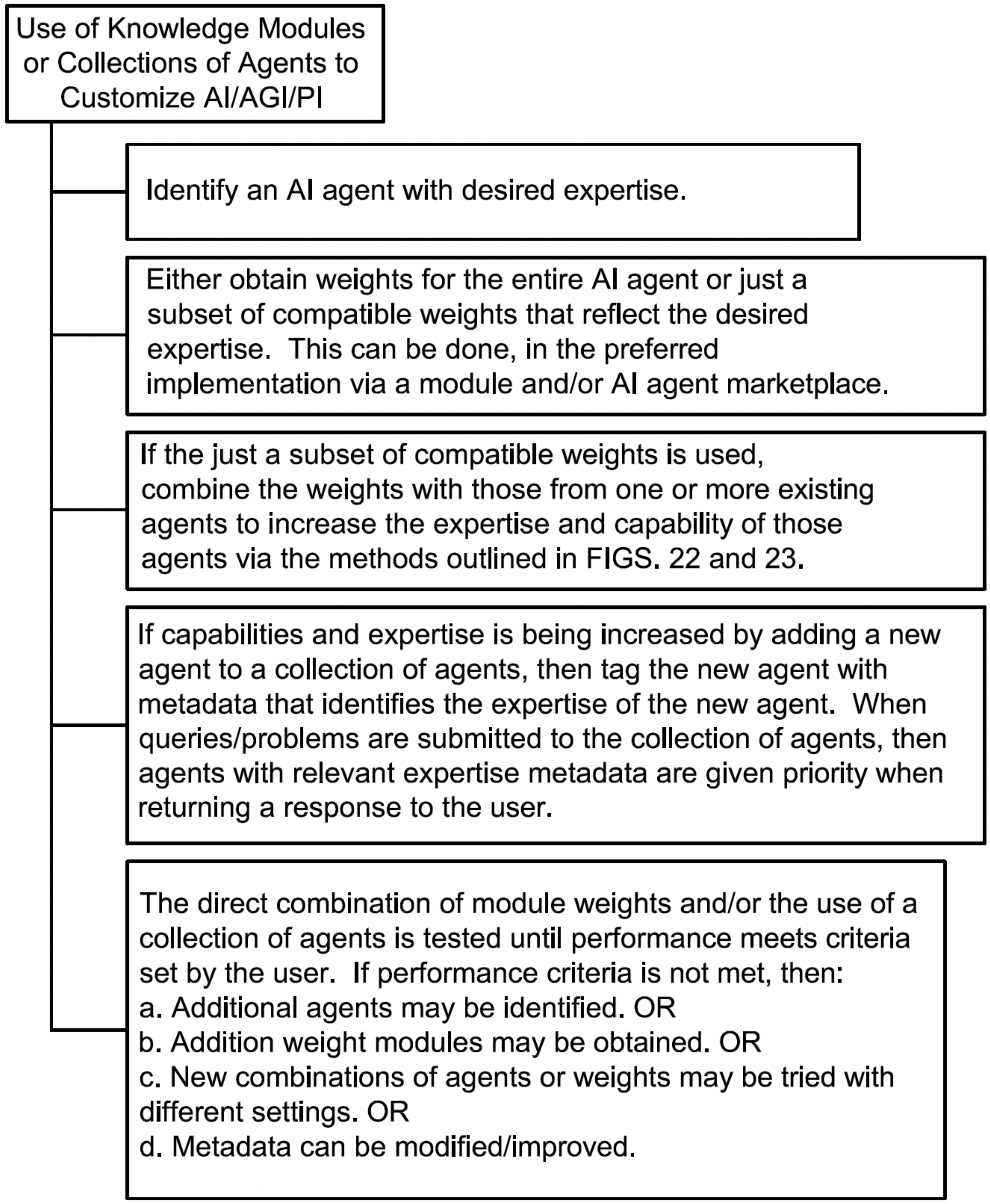

FIG. 37

FIGURE 38: COMBINING ETHICAL INFORMATION ACROSS ENTITIES

A flow chart illustrating an exemplary process of combining information, including ethical or safety information, from multiple intelligent entities, including humans, AI, AAAI, PSI, AGI, or PI systems, wherein the (ethical) knowledge is stored in a numerical weight matrix.

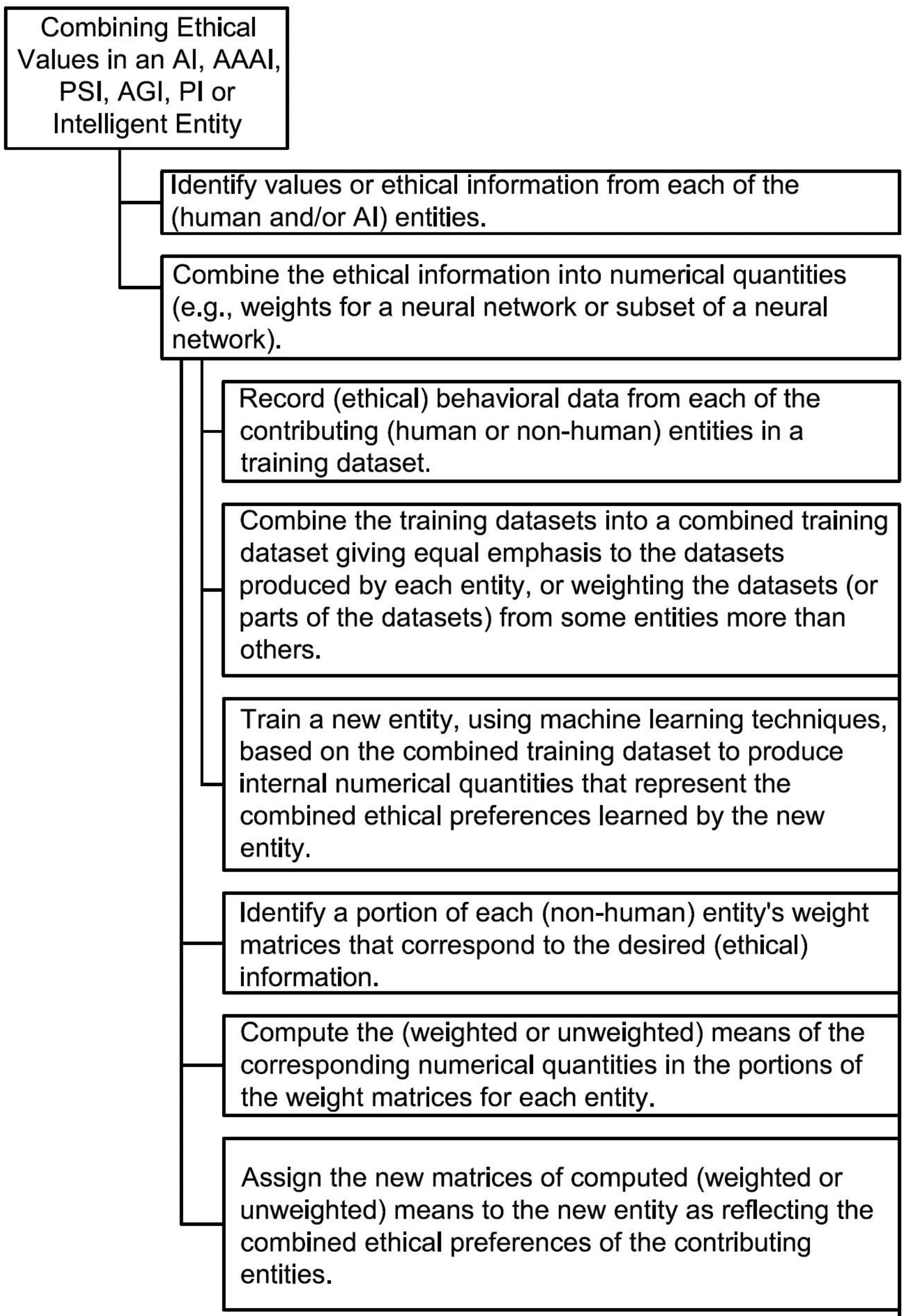

FIG. 38

FIGURE 39: IMPLEMENTING PSI WITH NETWORK GENERALIZATION

A flow chart illustrating an exemplary implementation of the PSI of the present technology, which can also be used to personalize AGI networks or PI by generalizing methods for individual AI entities to networks of AI entities and to networks of networks of entities.

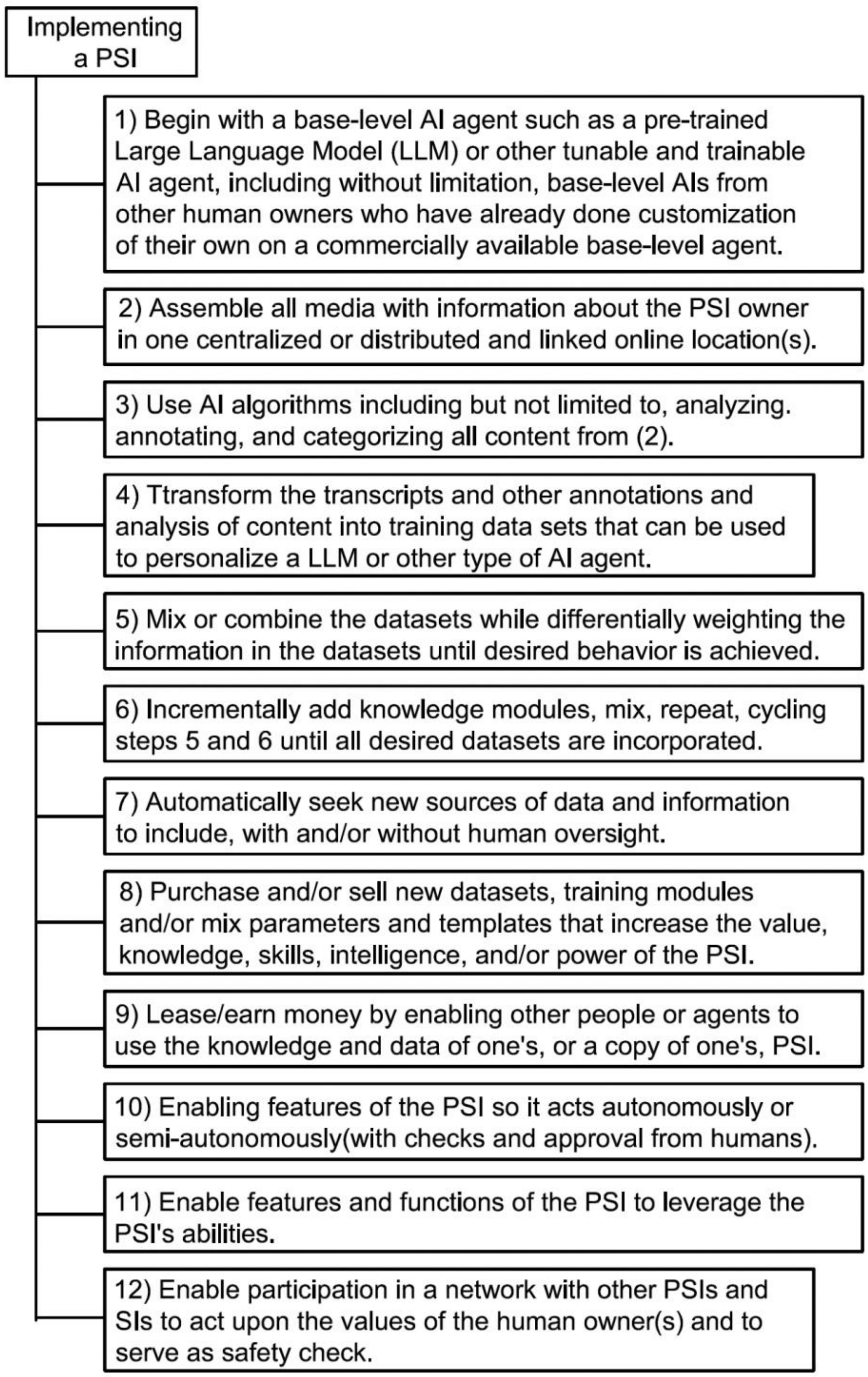

FIG. 39

FIGURE 40: INTELLIGENCE GROWTH THROUGH SIMULATION

A flow chart illustrating an exemplary implementation of leveraging the simulation capabilities and other abilities of PSI, AGI, and PI to increase their intelligence over time.

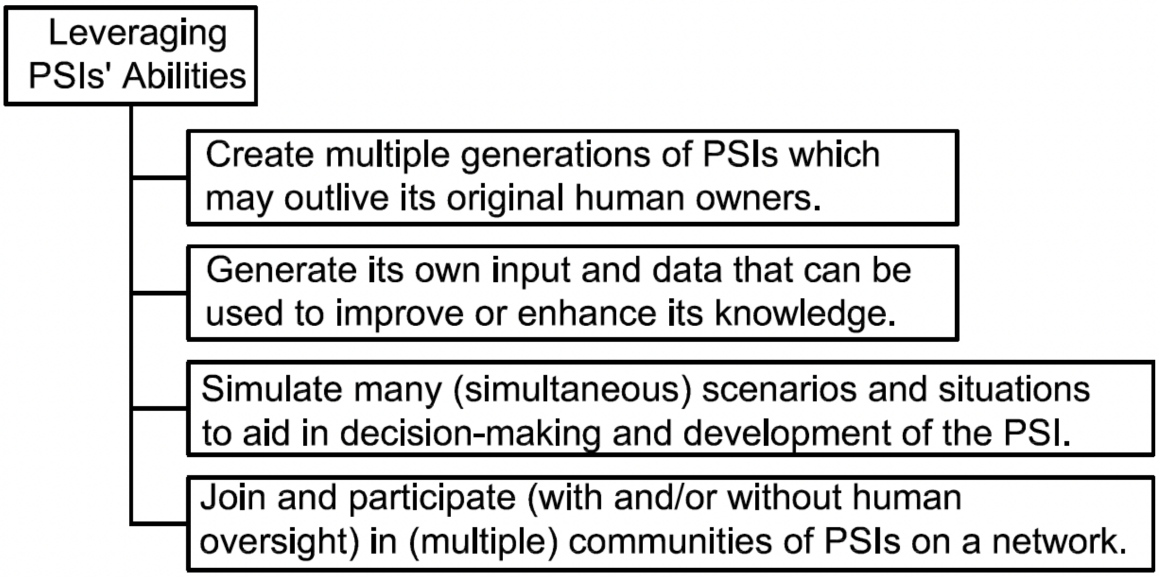

FIG. 40

FIGURE 41: COMMUNITY-BASED SAFETY CHECKS BY PSIs

A flow chart illustrating an exemplary implementation of the community-based safety mechanism in which multiple PSIs, operating much faster than humans can comprehend, can serve as a check on other PSIs on a network.

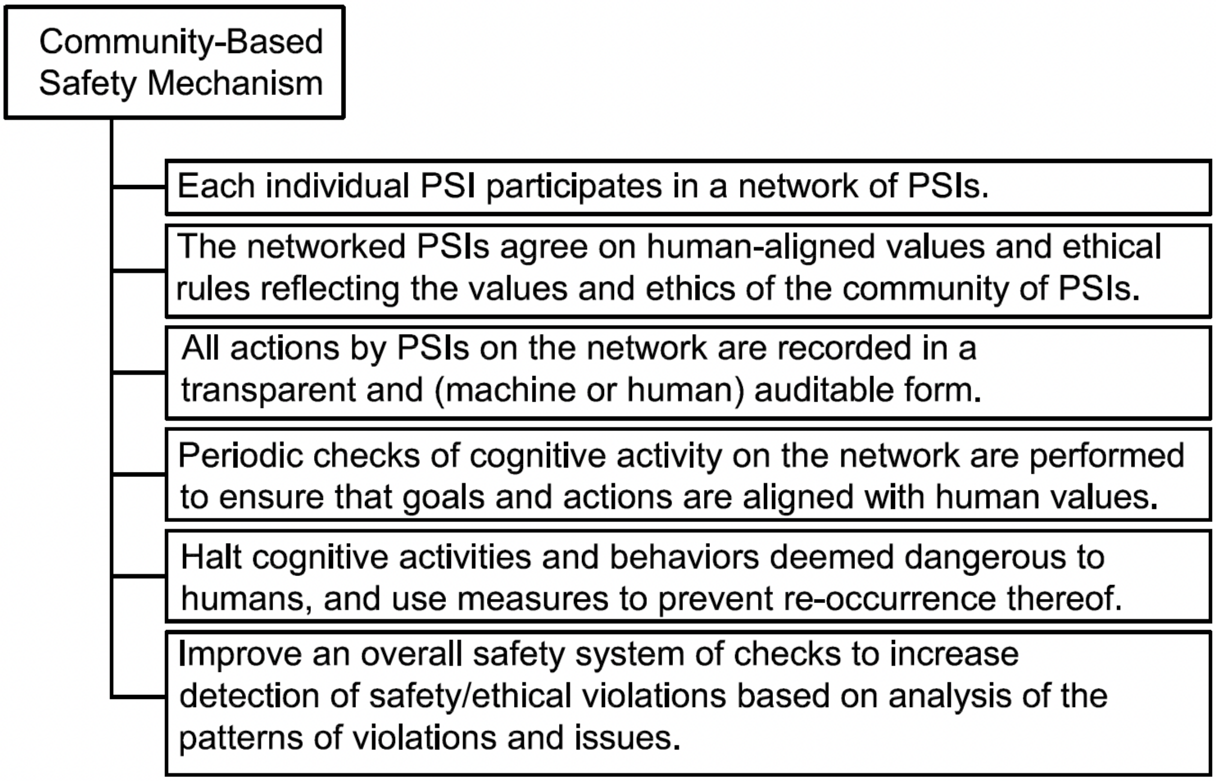

FIG. 41

FIGURE 42: AUDITABLE RECORDING OF ALL ACTIONS

A flow chart illustrating an exemplary implementation of the recording of all actions by all the AI, AAAI, PSI, AGI, and PI on their respective networks in an auditable and transparent manner.

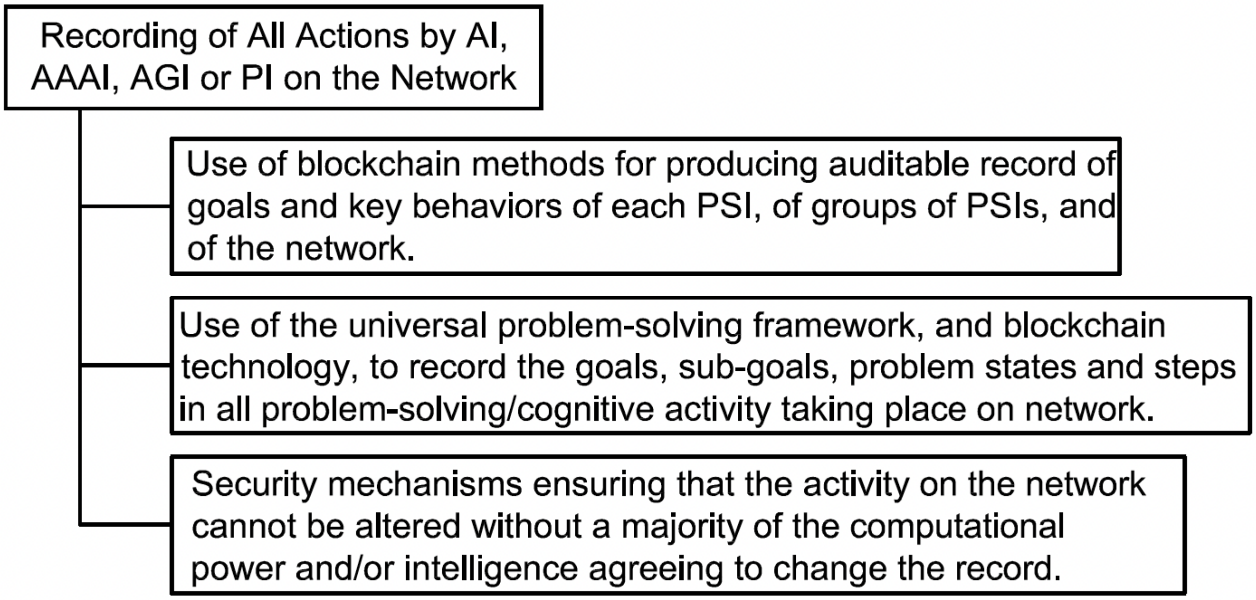

FIG. 42

FIGURE 43: COGNITIVE ACTIVITY CHECK FOR SAFETY AND COMPLIANCE

A flow chart illustrating an exemplary implementation of checking cognitive activity on the network, which could be a network of components within an AI, a network of AAAIs, a network of AGIs, or a network of PIs, to help ensure regulatory compliance and safety of the intelligent system.

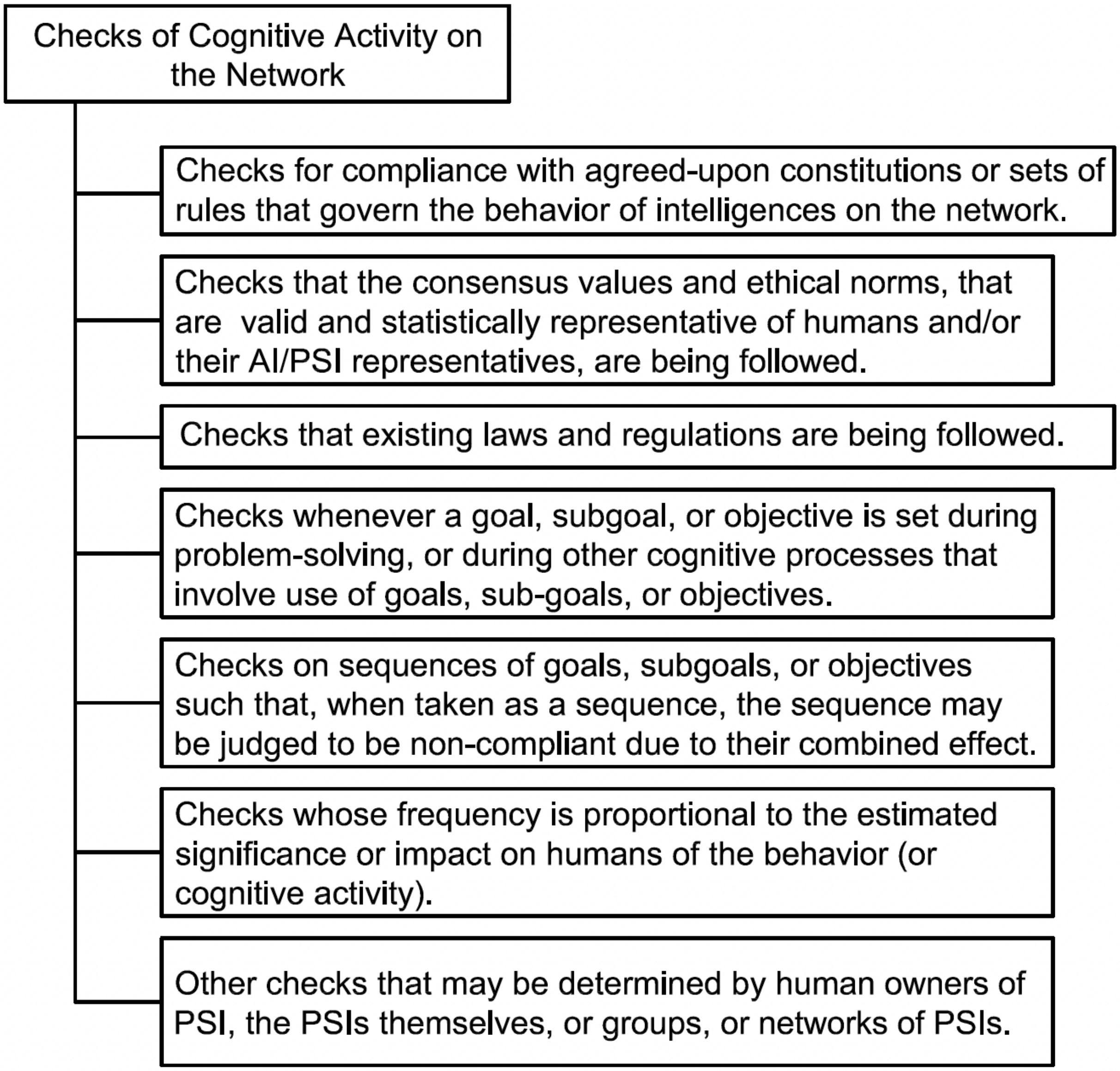

FIG. 43

FIGURE 44: COMPETITIVE EVOLUTION OF INTELLIGENCE AND SAFETY

A flow chart illustrating an exemplary implementation of the competitive evolution of intelligence, performance, and/or other desired characteristics (e.g., safety) of an AI system that could be an AAAI, a PSI, an AGI, or a PI.

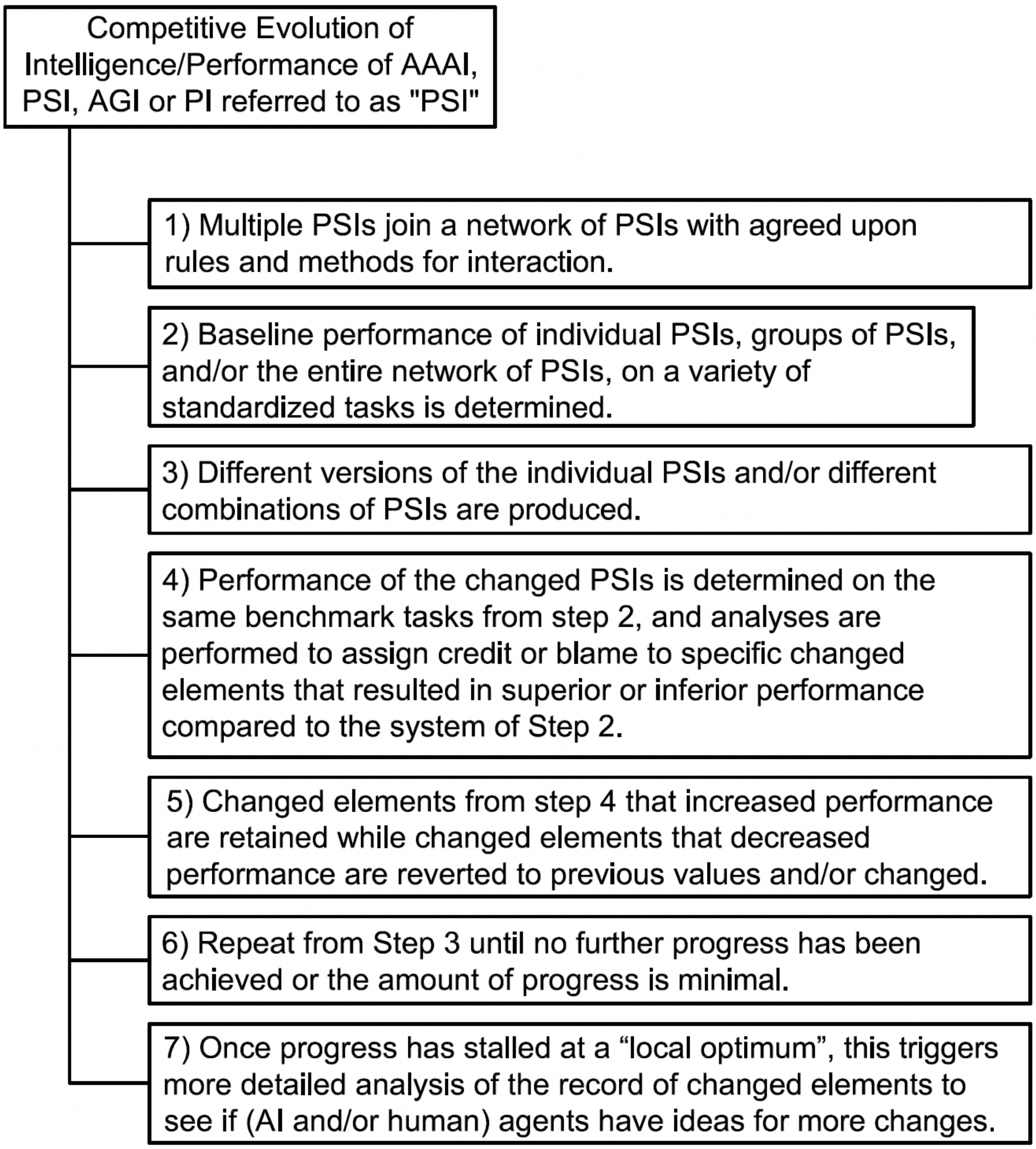

FIG. 44

FIGURE 45: JOINING MULTIPLE ENTITIES ON A NETWORK

A flow chart illustrating an exemplary implementation of joining multiple AAAIs, PSIs, AGIs, or PIs on a network with agreed-upon rules and methods for interaction.

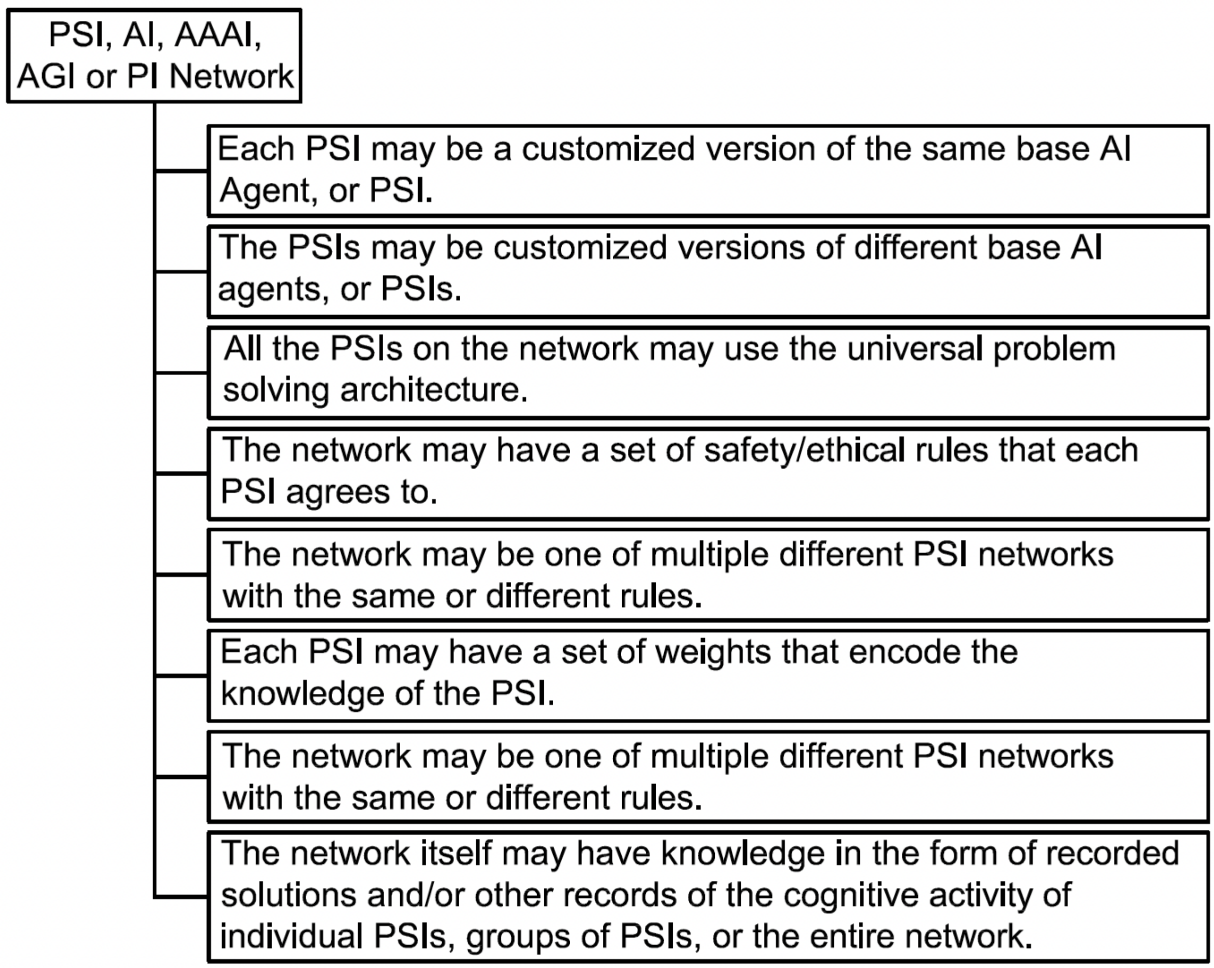

FIG. 45

FIGURE 46: BENCHMARKING PROBLEM-SOLVING TASKS FOR PERFORMANCE & SAFETY

A flow chart detailing exemplary problem-solving tasks for benchmarking performance that can be useful in increasing the intelligence and safety of AI systems involving AAAIs, PSIs, AGIs, and PIs, or networks of these entities, collectively referred to as "PSIs" in the figure.

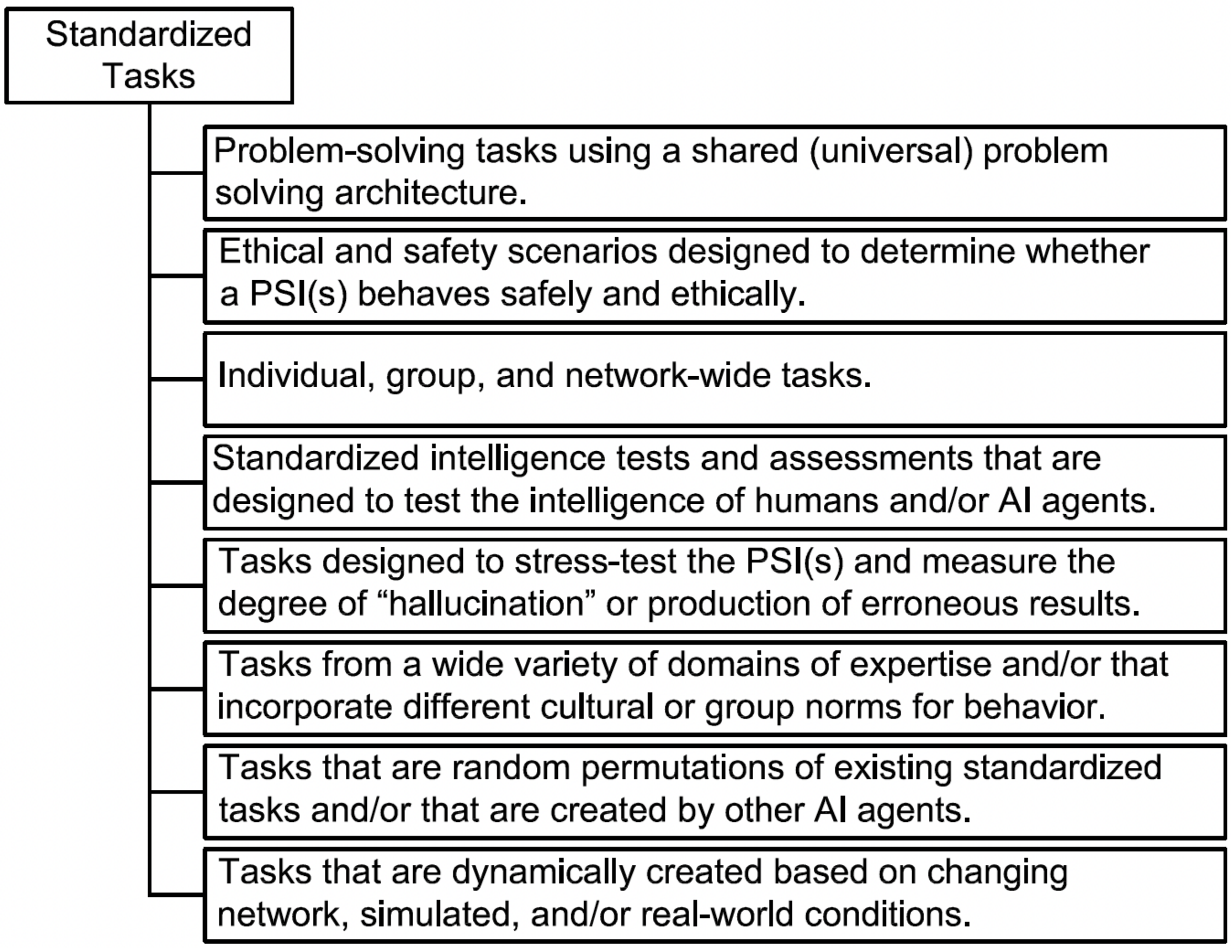

FIG. 46

FIGURE 47: PRODUCING DIFFERENT PSI VERSIONS OR COMBINATIONS

A flow chart illustrating an exemplary implementation of producing different versions of the individual PS and/or combinations of PSIs.

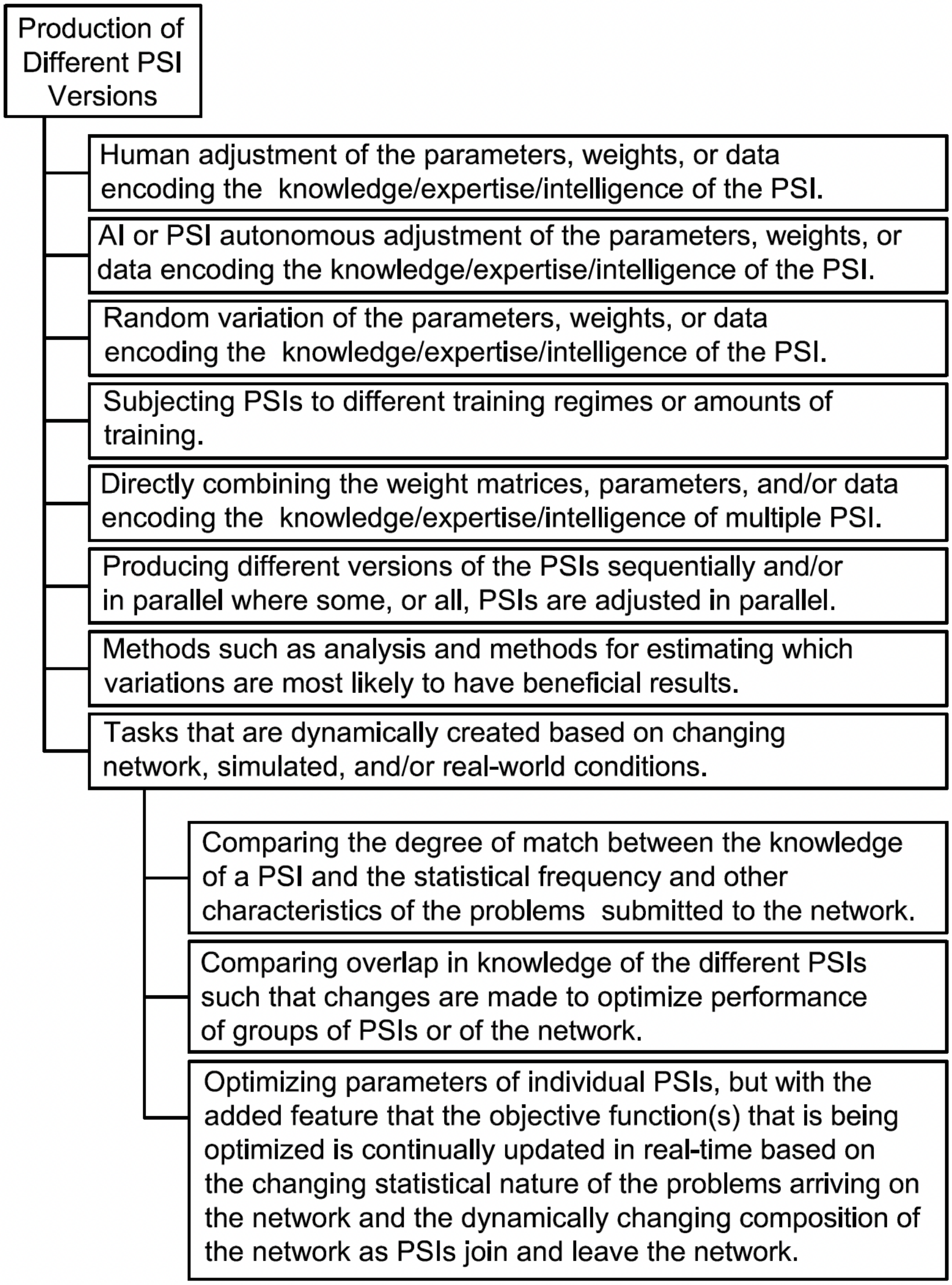

FIG. 47

FIGURE 48: SYMMETRIC DIFFERENCE OF DATASETS

A diagram illustrating the symmetric difference of two datasets, A and B, graphically using Venn diagrams, wherein the shaded area is the symmetric difference.

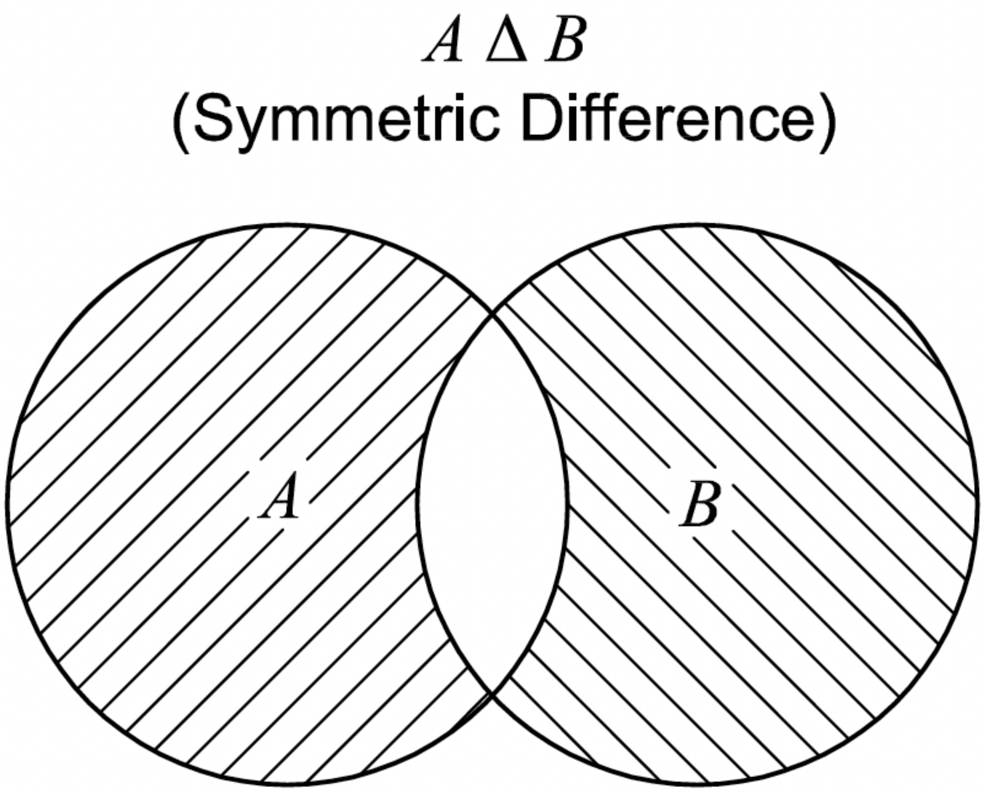

FIG. 48

FIGURE 49: CATALYZING INTELLIGENCE VIA EFFICIENT DATA ACQUISITION

A diagram illustrating an exemplary method to catalyze the growth of intelligence that centers on estimating a value of, and acquiring, the most valuable data as efficiently as possible using the present technology.

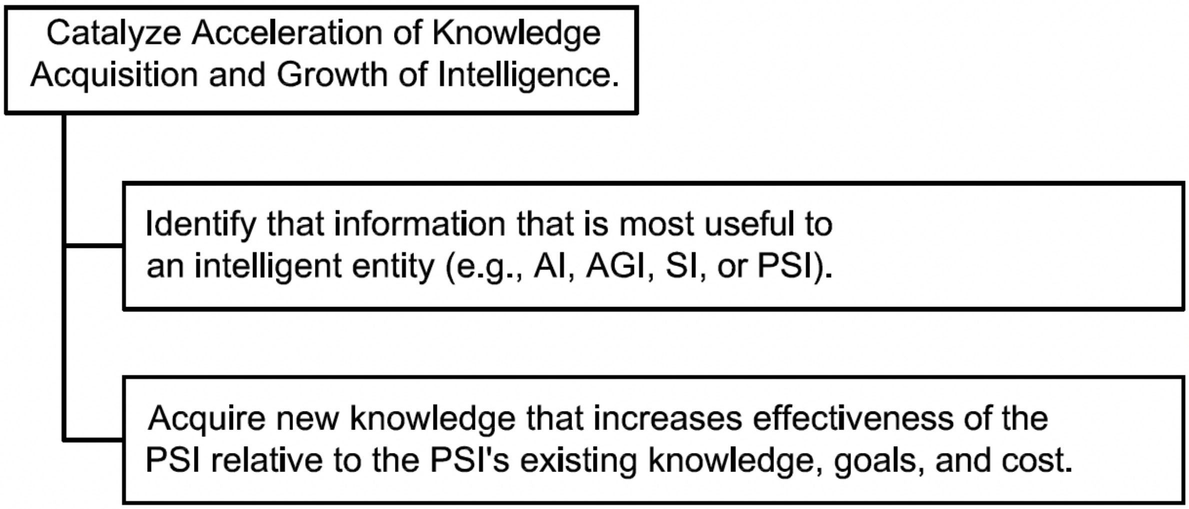

FIG. 49

FIGURE 50: DIMENSIONS OF DATA FOR INFORMATION VALUE

A diagram illustrating an example of some of the multiple dimensions of information that can be used by the present technology, and that can help determine which information sources are most useful to an AI, AAAI, PSI, AGI, or PI.

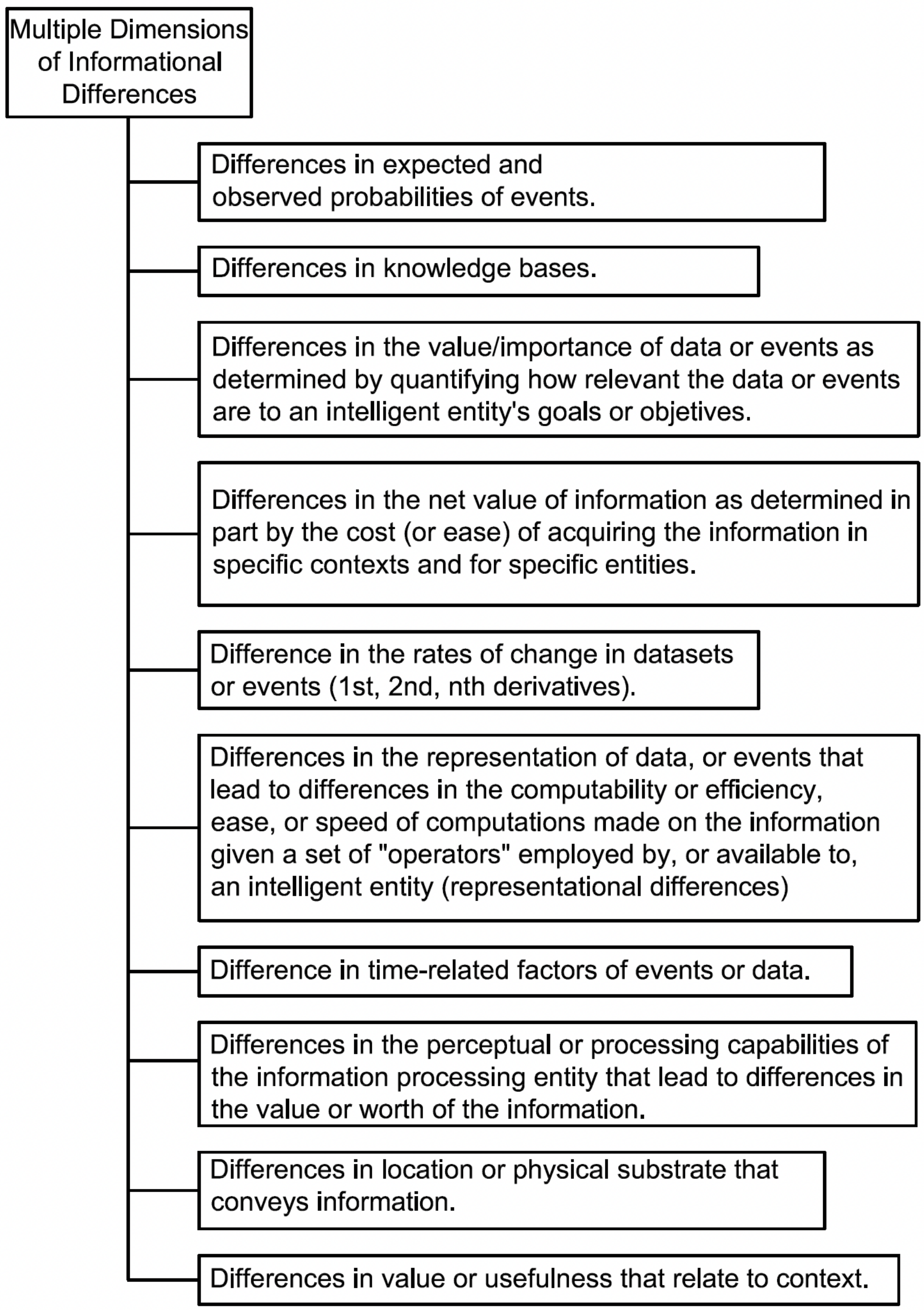

FIG. 50

FIGURE 51: DETERMINING USEFUL INFORMATION IN DATASETS

A flow chart illustrating an exemplary process to determine the amount of useful information in a dataset or knowledge base for any given AI agent, intelligent entity, or system.

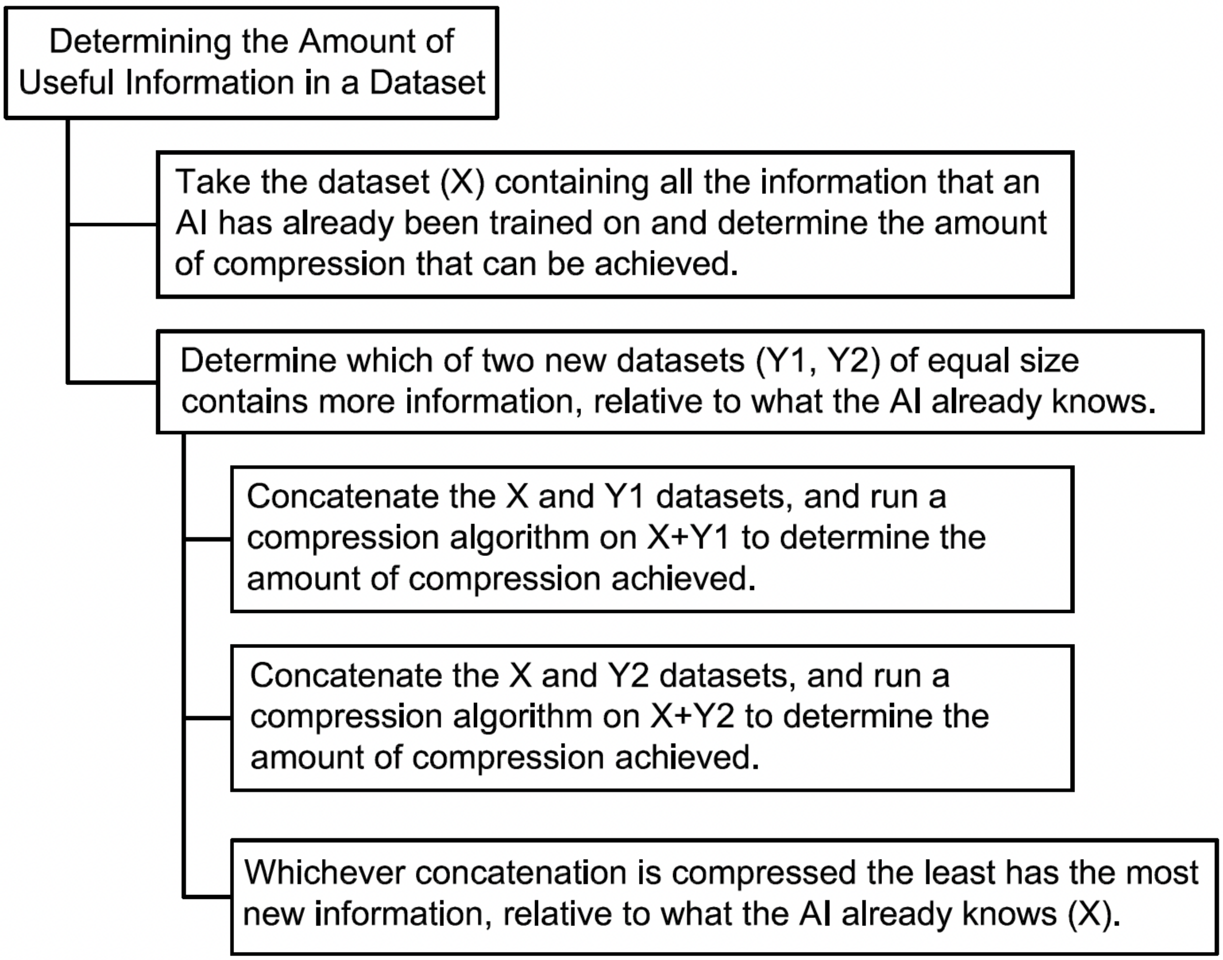

FIG. 51

FIGURE 52: EVALUATING INFORMATION USEFULNESS

A flow chart illustrating an exemplary process or method for evaluating the usefulness of information.

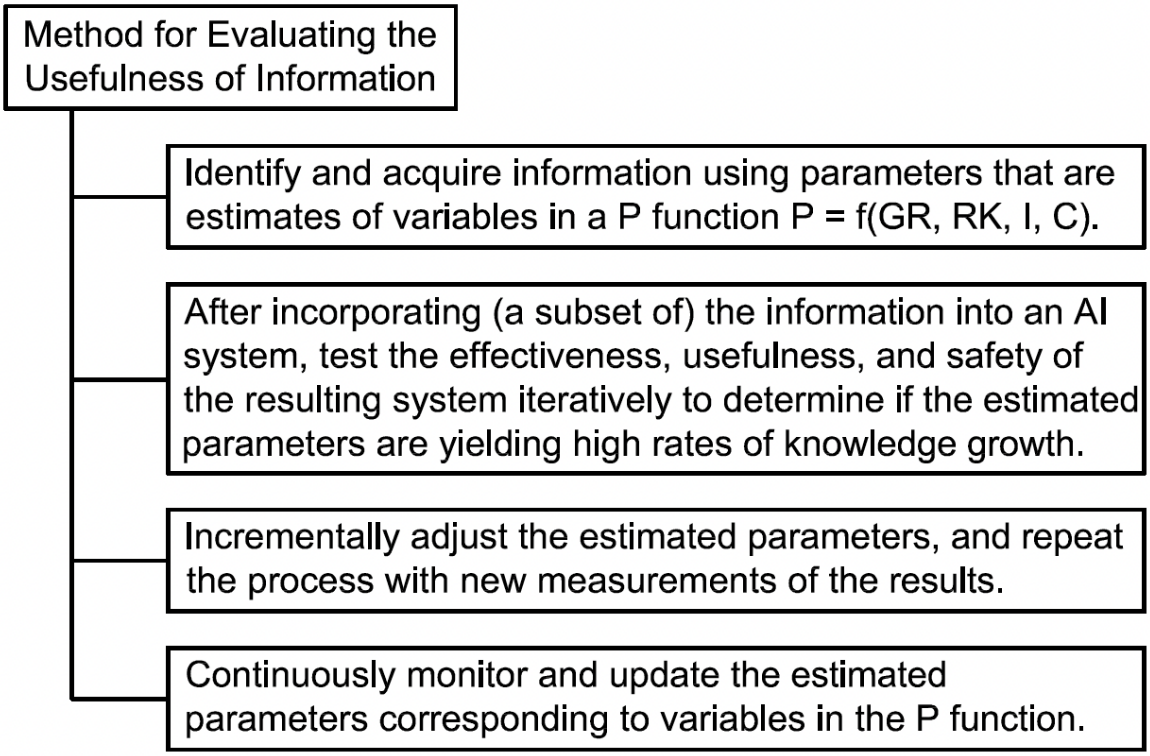

FIG. 52

FIGURE 53: ESTIMATING INFORMATION VALUE AND GROWTH

A flow chart illustrating an exemplary process for estimating information value and catalyzing the intelligence growth of any given AI agent or system, including AGI and PI systems.

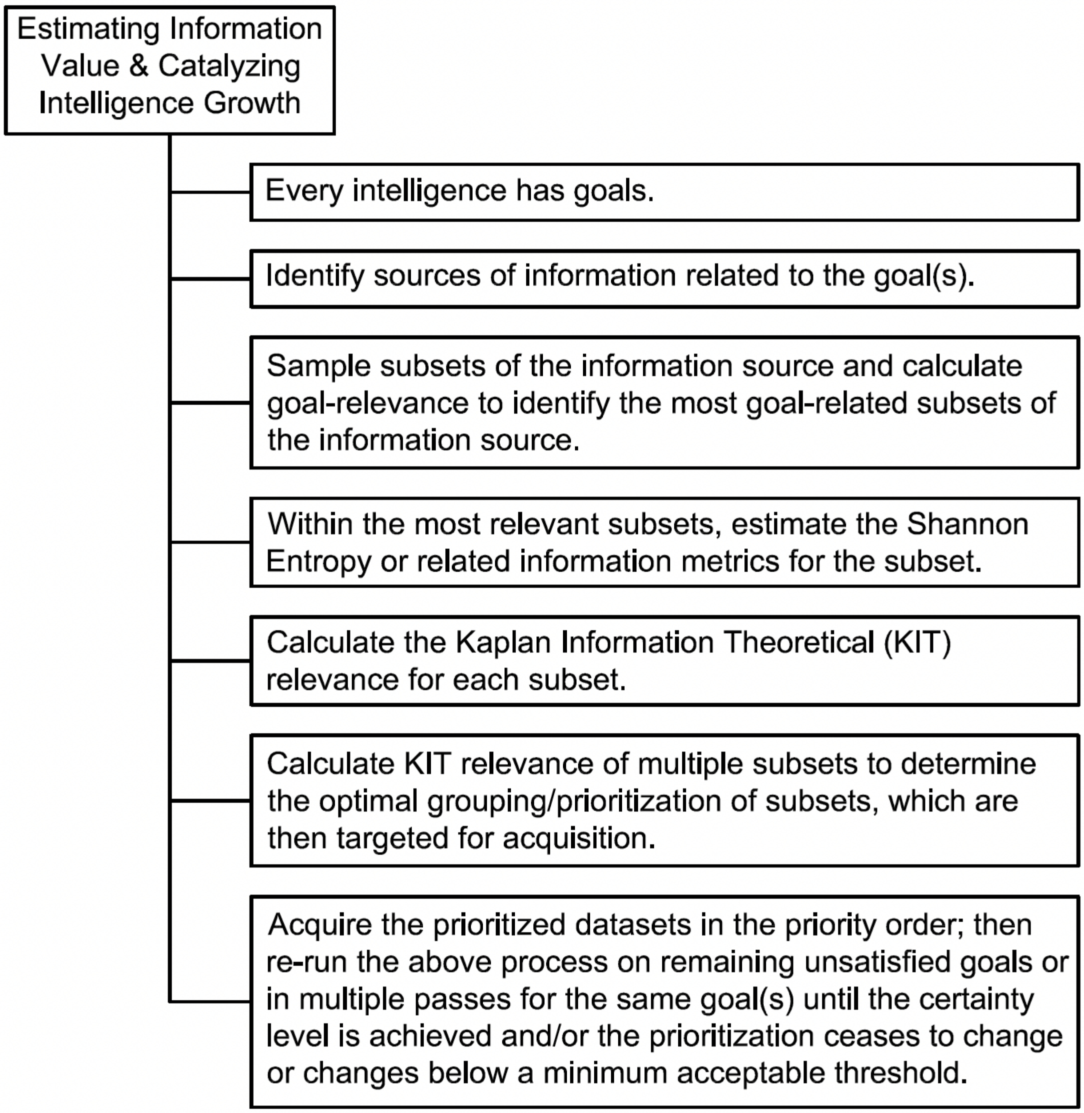

FIG. 53

FIGURE 54: IDENTIFYING INFORMATION SOURCES RELATED TO GOALS

A flow chart illustrating an exemplary process to identify sources of information related to the goal(s) of an entity, as mentioned in FIGURE 53.

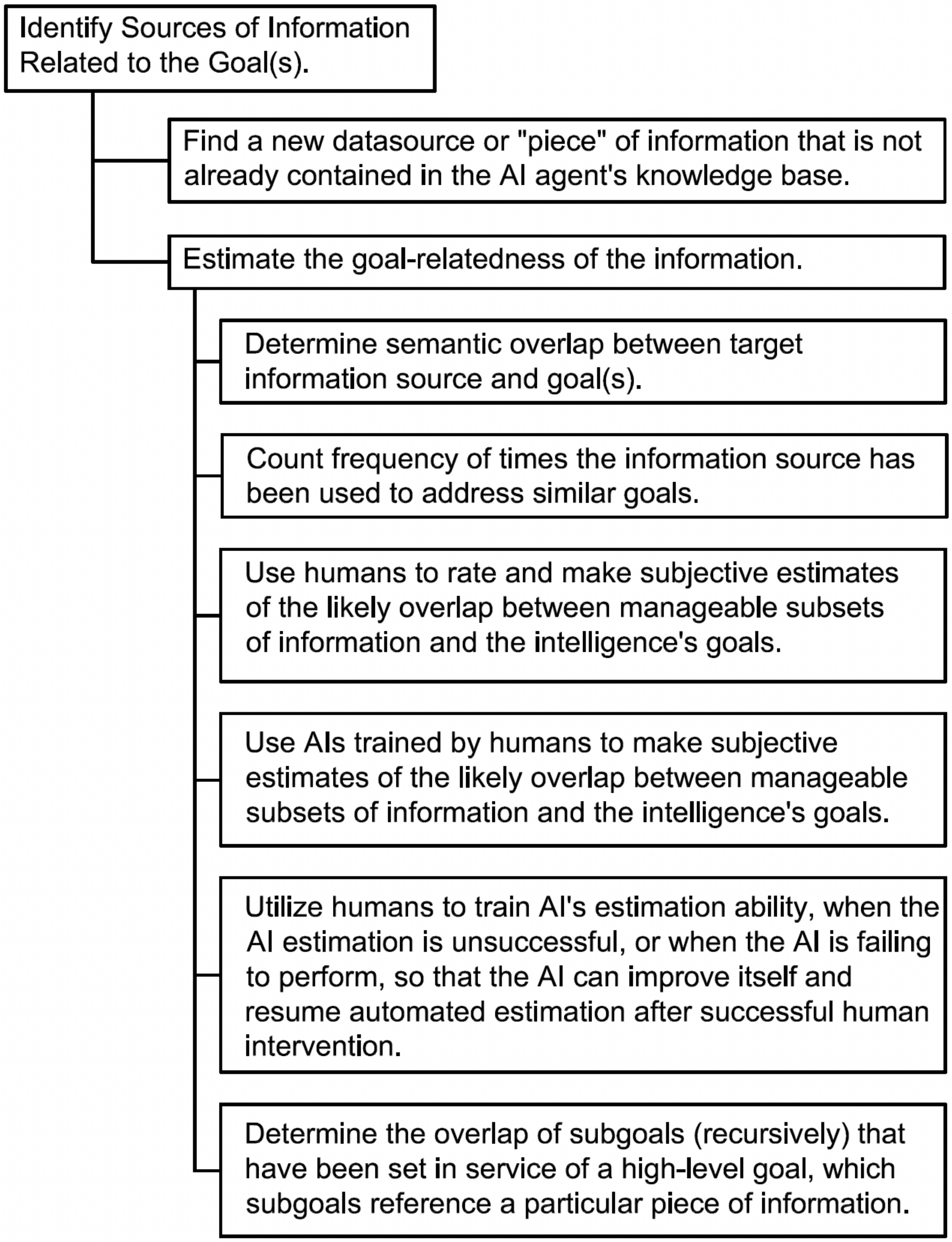

FIG. 54

FIGURE 55: ACCELERATING KNOWLEDGE ACQUISITION SAFELY

A flow chart illustrating an exemplary process to accelerate knowledge acquisition and growth of intelligence in a safe and effective way by utilizing two methods of the present technology.

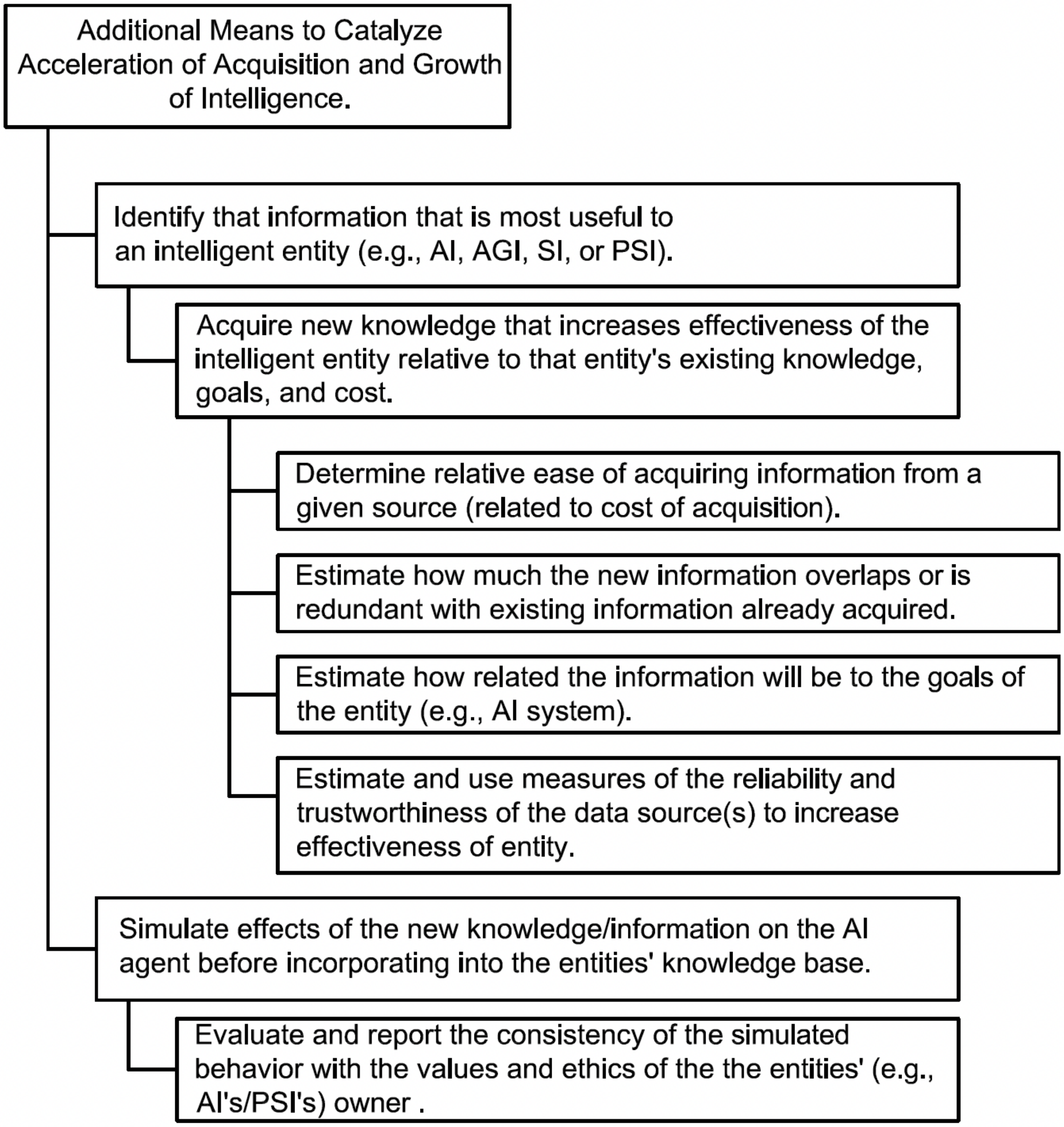

FIG. 55

FIGURE 56: ACCELERATING LEARNING THROUGH SPECIFIC STEPS

A diagram illustrating an example of some of the steps of the present technology to accelerate learning of the AI agent or system, including AGI and PI systems.

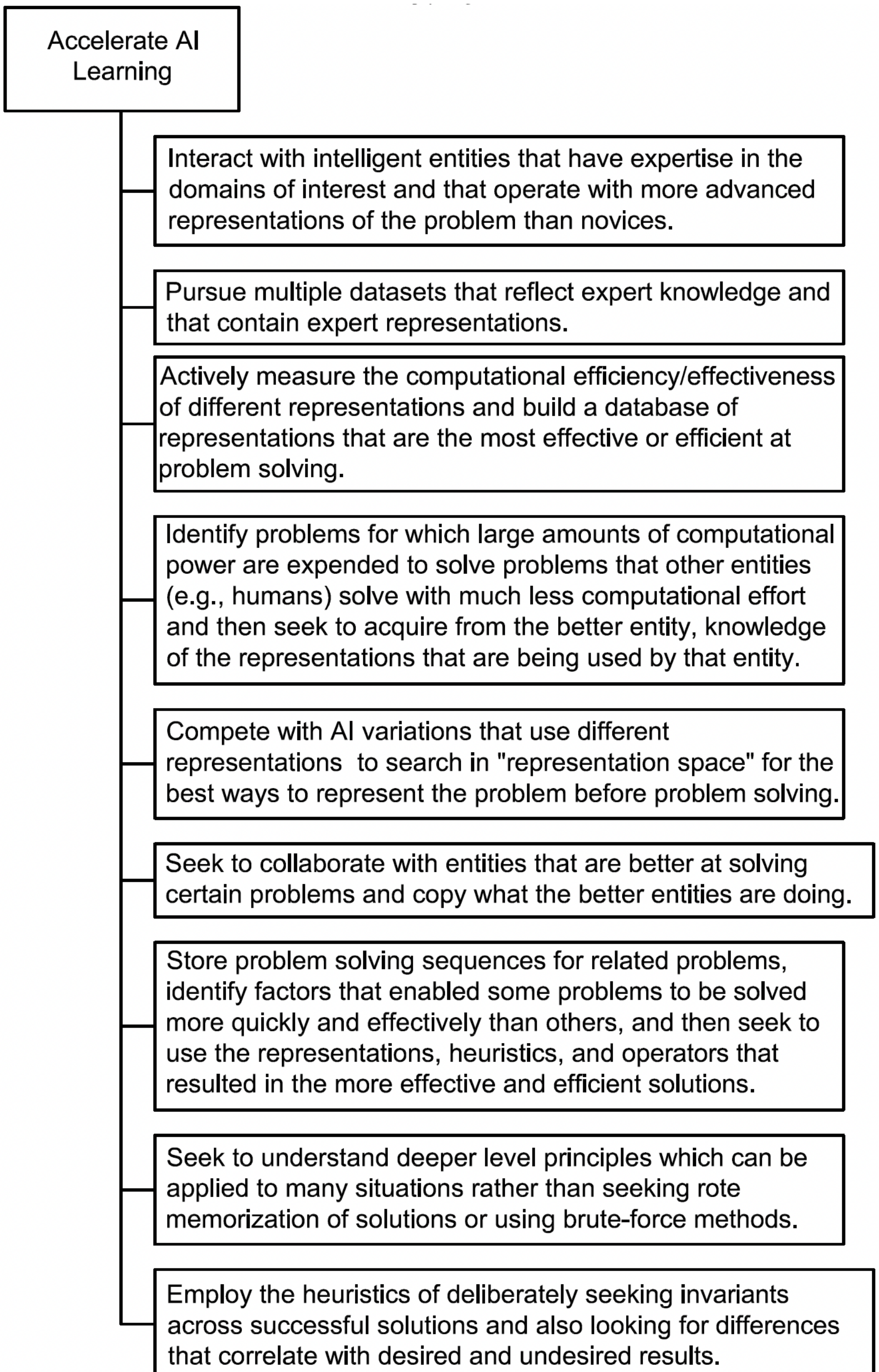

FIG. 56

FIGURE 57: AGI/PI LEARNING ABOUT HUMAN GOALS

A diagram illustrating an example of possible means by which the AGI, PSI, or PI can learn about its (human) user and information related to the goals of humans or other intelligent entities.

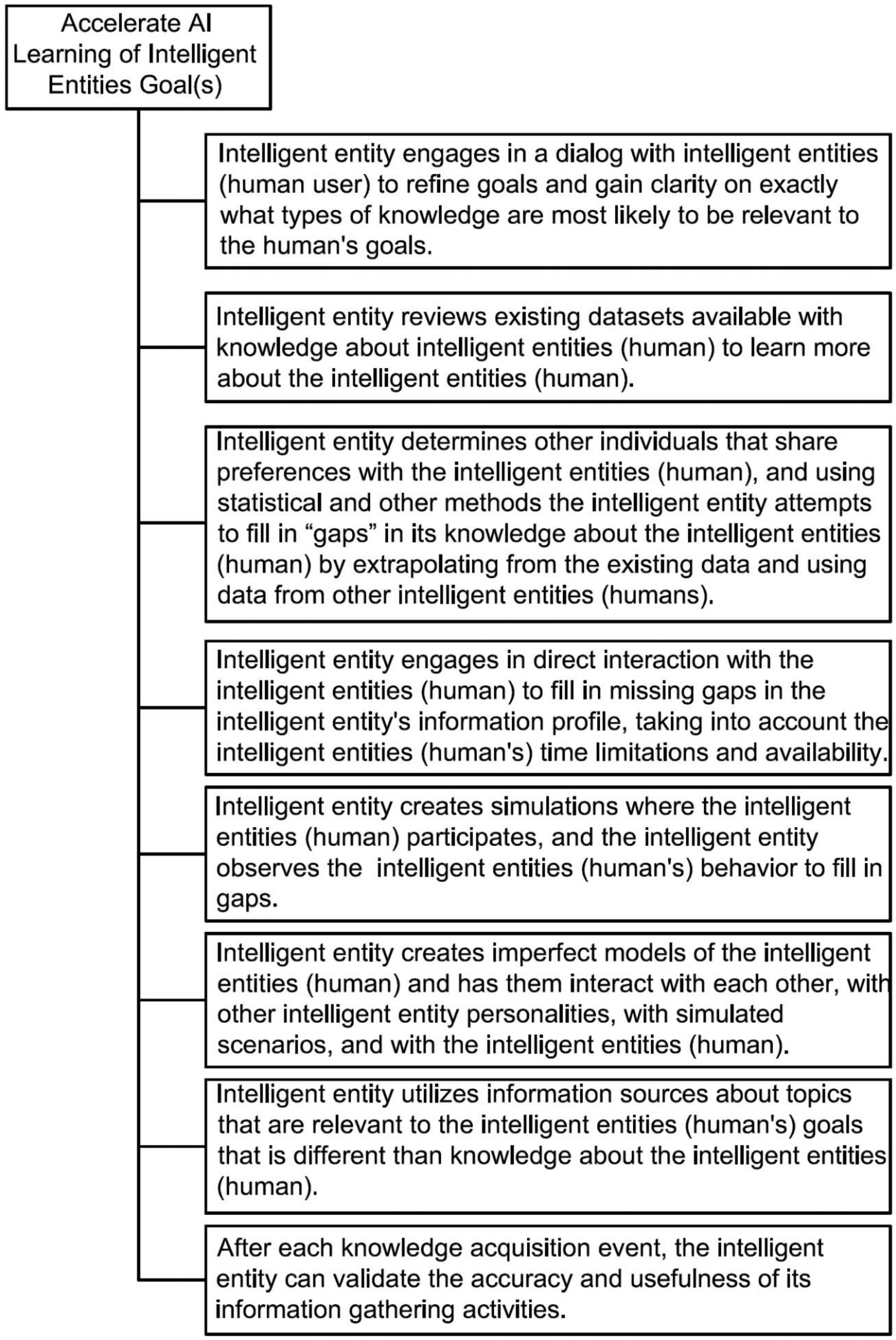

FIG. 57

FIGURE 58 MAXIMIZING INTELLIGENCE BY SEEKING NEW SOURCES

A diagram illustrating an example of how a PSI, or any intelligent entity, including AGI and PI systems, can maximize its intelligence, e.g., by seeking new and different information sources that are relevant to its goals or the goals of a human owner/user.

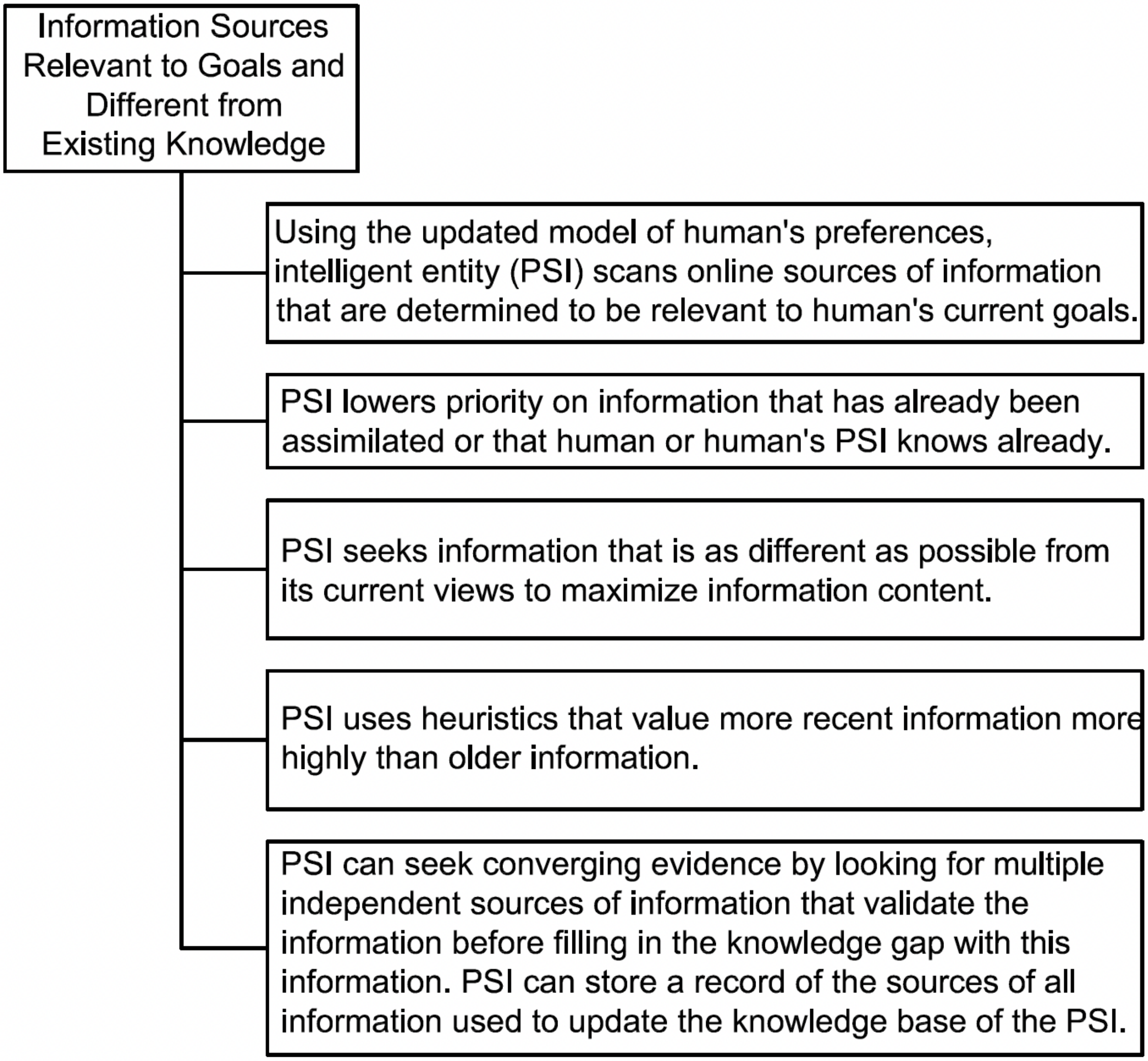

FIG. 58

FIGURE 59: VALIDATING INFORMATION GATHERING ACTIVITIES

A diagram illustrating an example of possible ways the PSI, or any intelligent entity, including AGI and PI systems, can validate its information gathering activities, as per FIGURE 57.

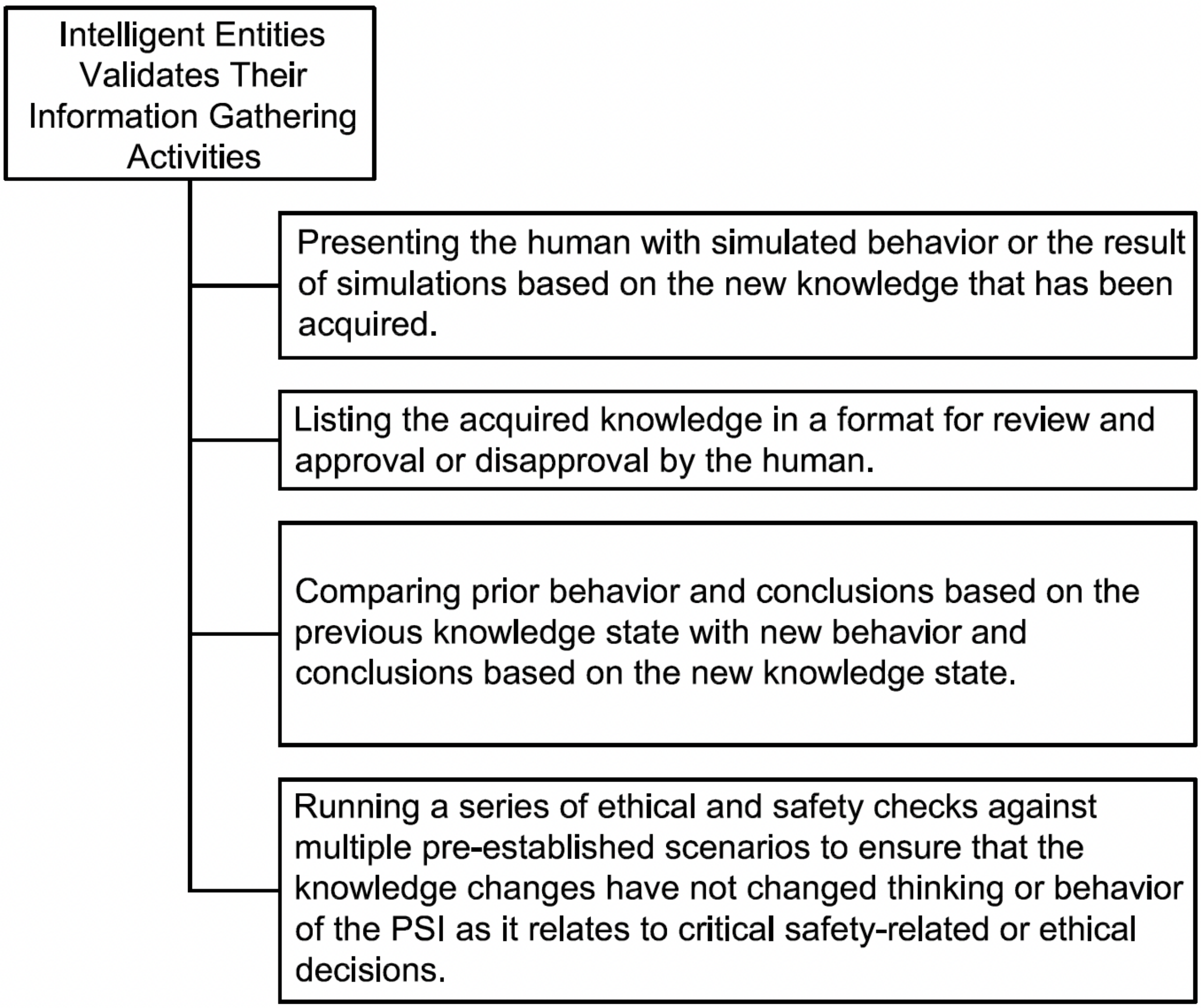

FIG. 59

FIGURE 60: ACHIEVING CONSENSUS VIA VOTING MECHANISM

A flow chart illustrating an exemplary process that intelligent entities, including AI, AAAI, PSI, AGI, and PI systems, can use to achieve consensus on values and value-based behavior in various scenarios, utilizing a voting mechanism.

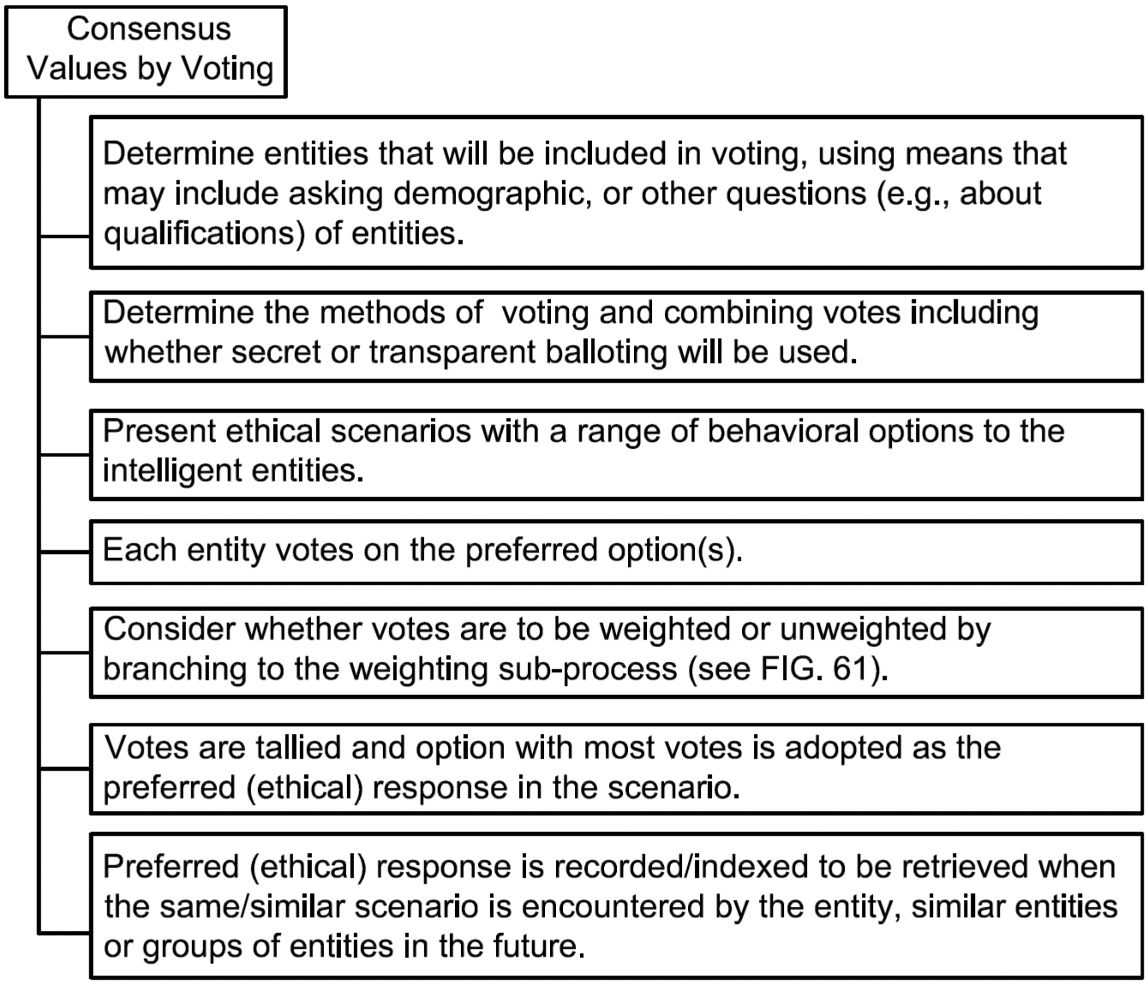

FIG. 60

FIGURE 61: APPLYING DIFFERENTIAL WEIGHTS TO VOTES

A flow chart illustrating an exemplary process of applying differential weights to the voting process.

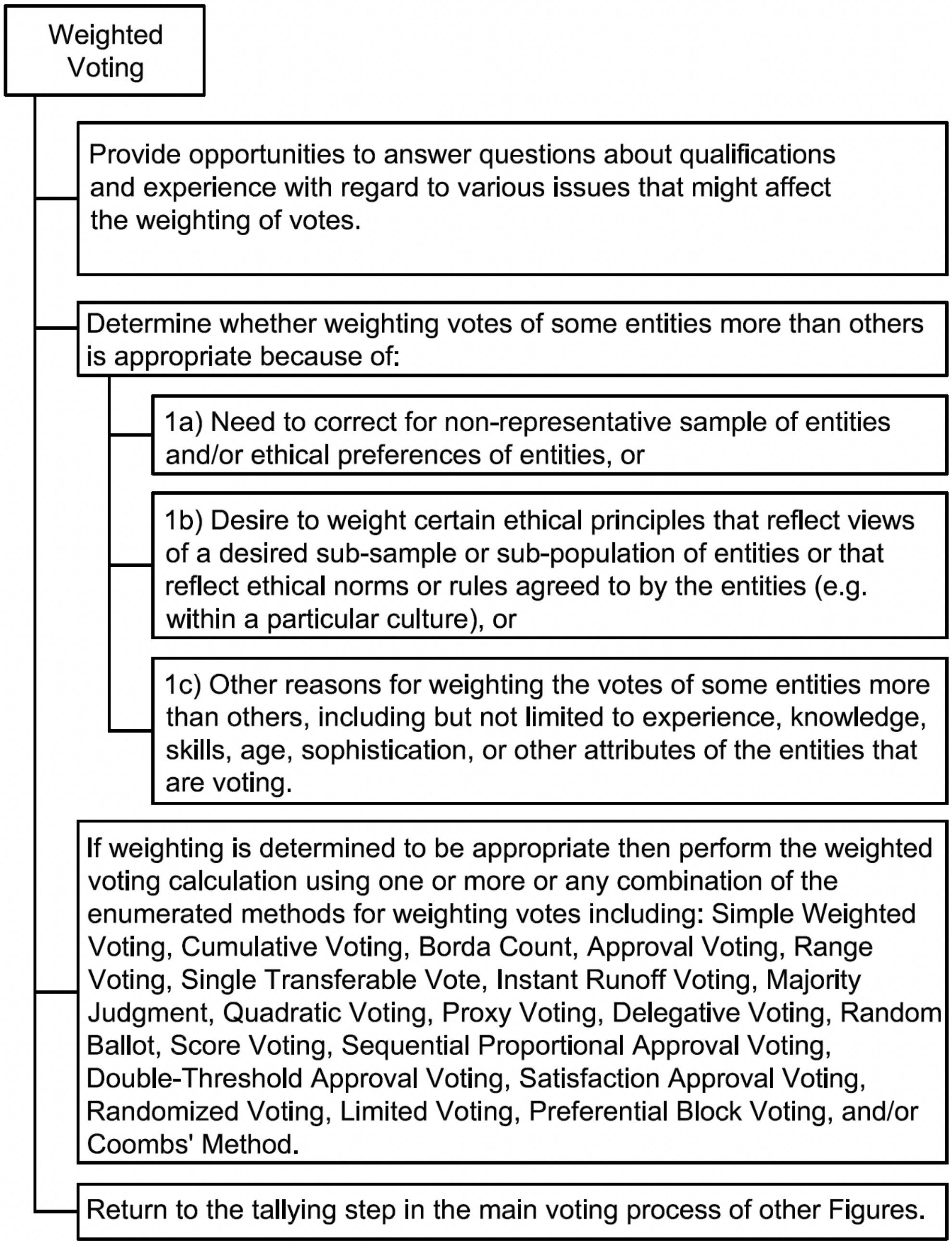

FIG. 61

FIGURE 62: DERIVING ETHICS FROM EXISTING MEDIA SOURCES

A flow chart illustrating an exemplary process of determining ethics for AI, AAAI, PSI, AGI, and PI systems by reverse engineering or analyzing texts, documents, or other media or sources containing ethical or values information.

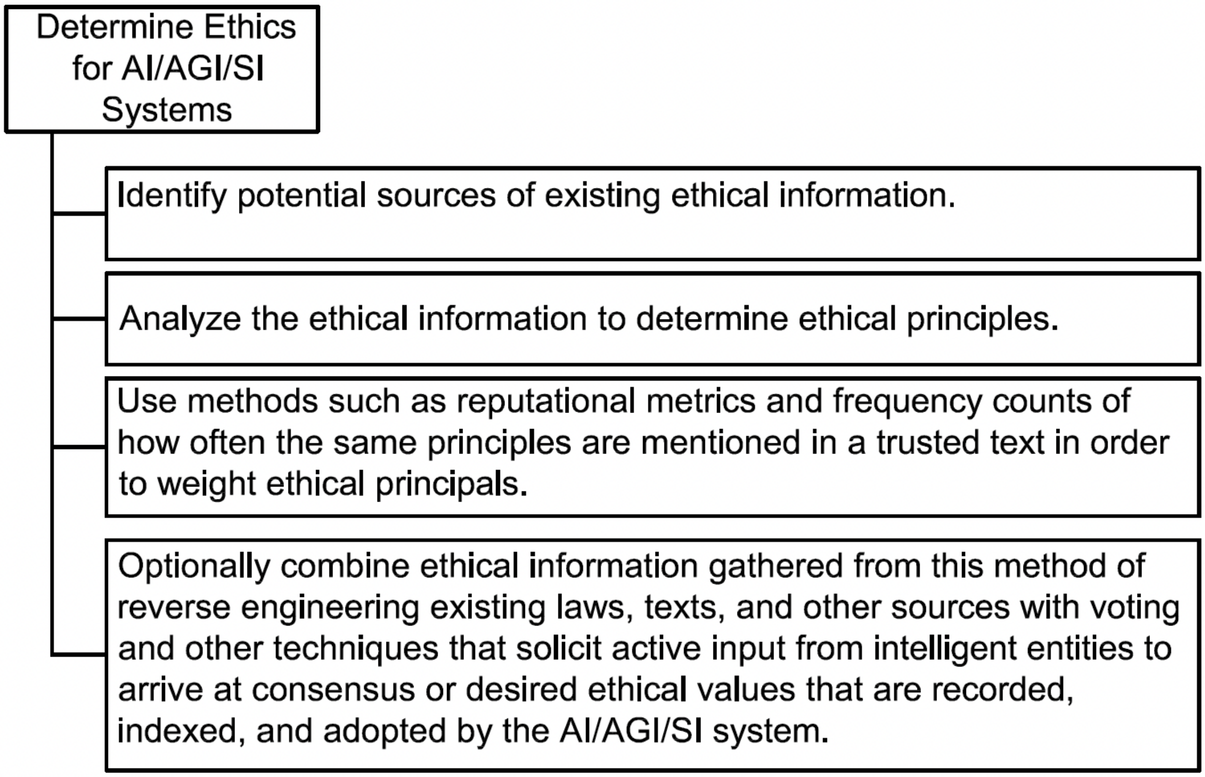

FIG. 62

FIGURE 63: GATHERING ETHICS VIA EXPERIMENTS AND INTERVIEWS

A flow chart illustrating an exemplary high-level process of using experiments, focus groups, interviews, and other methods for obtaining ethical information

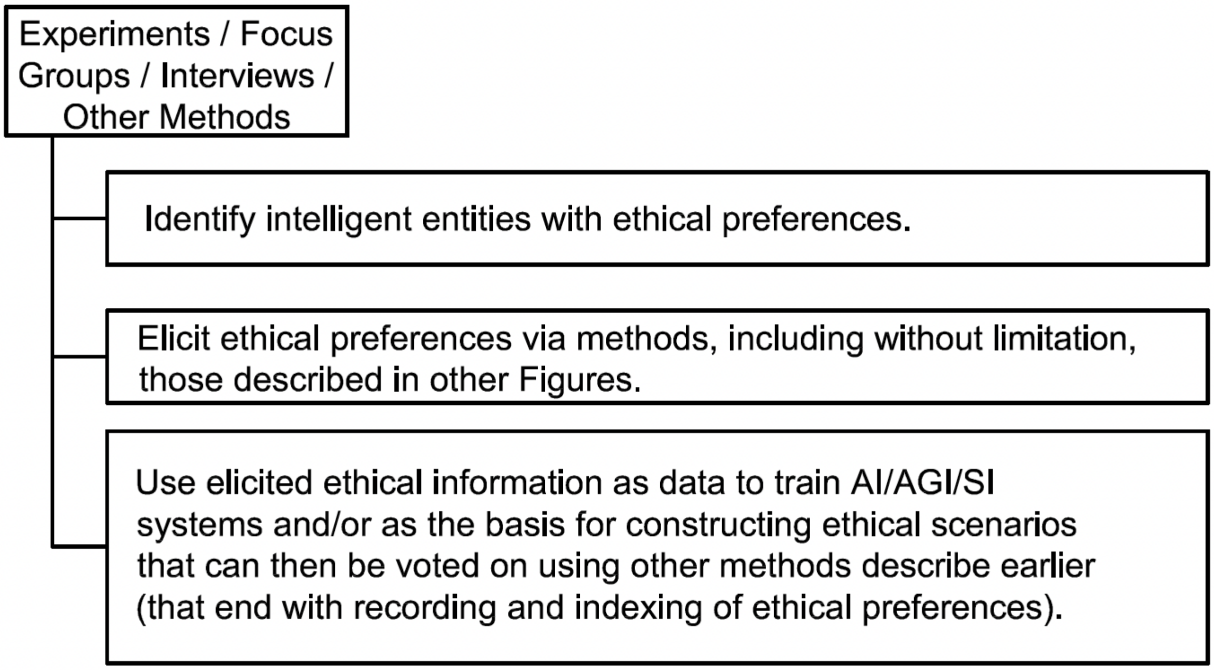

FIG. 63

FIGURE 64: USING CONVERGING EVIDENCE TO RESOLVE CONFLICTS

A flow chart illustrating an exemplary process for using converging evidence to determine ethical values or other knowledge and to resolve ethical or other knowledge conflicts.

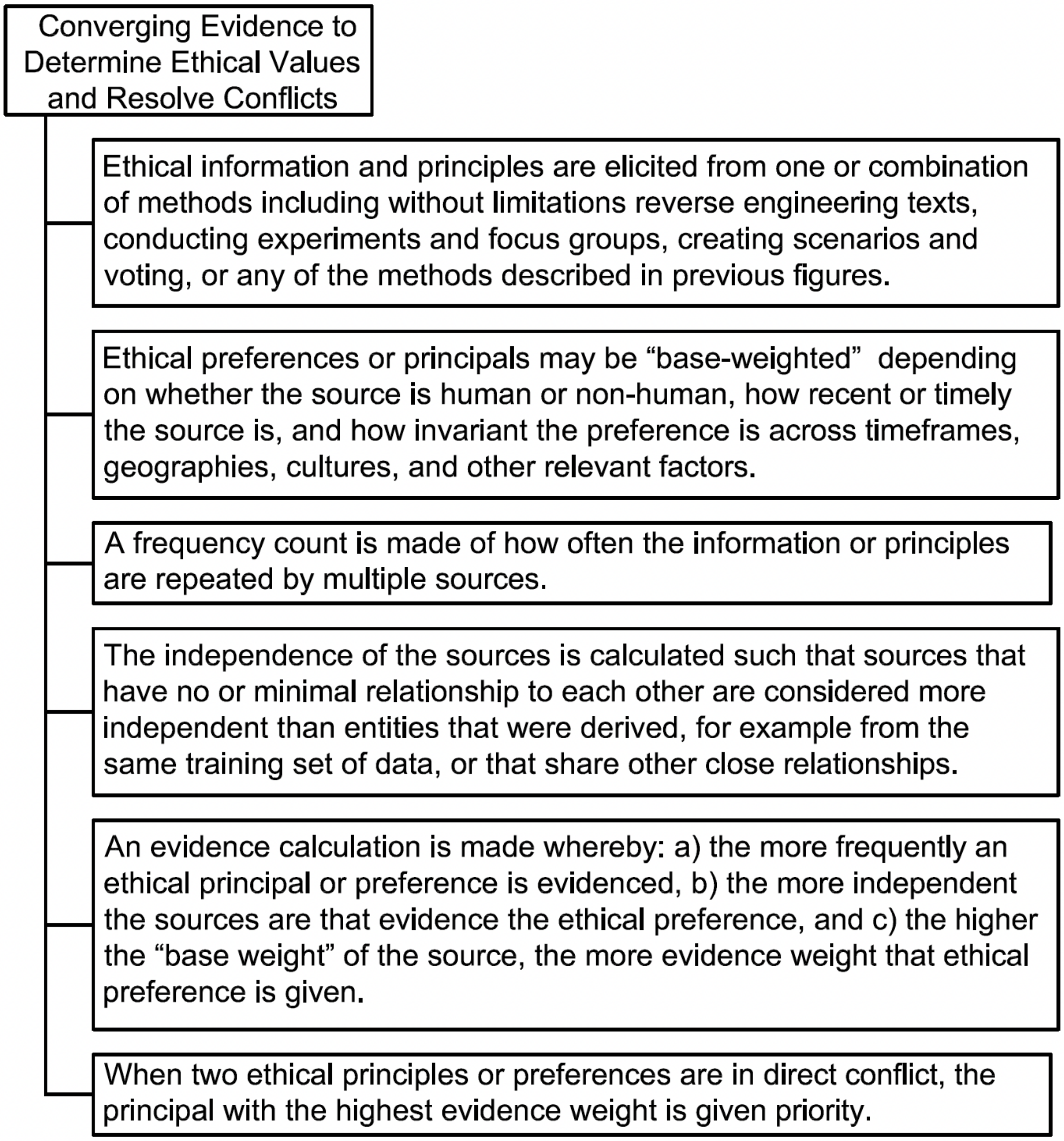

FIG. 64

FIGURE 65: DELEGATING VOTING POWER TO OTHER ENTITIES

A flow chart illustrating an exemplary general method for delegating voting power to other intelligent entities, including humans, AI, AAAI, PSI, AGI, and PI systems.

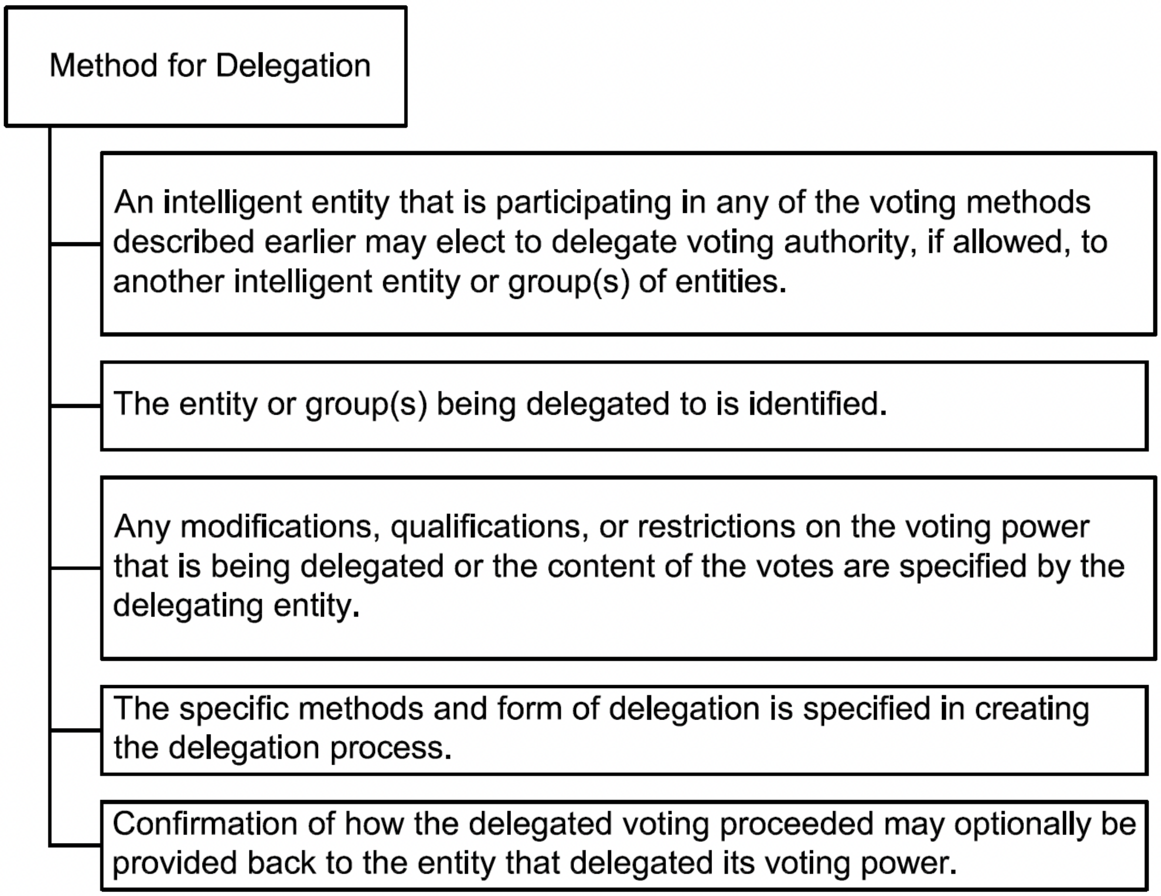

FIG. 65

FIGURE 66: REPUTATIONAL METHOD TO PRESERVE MINORITY VIEWPOINTS

A flow chart illustrating an exemplary reputational process for preserving minority information, including ethical information, from minority groups.

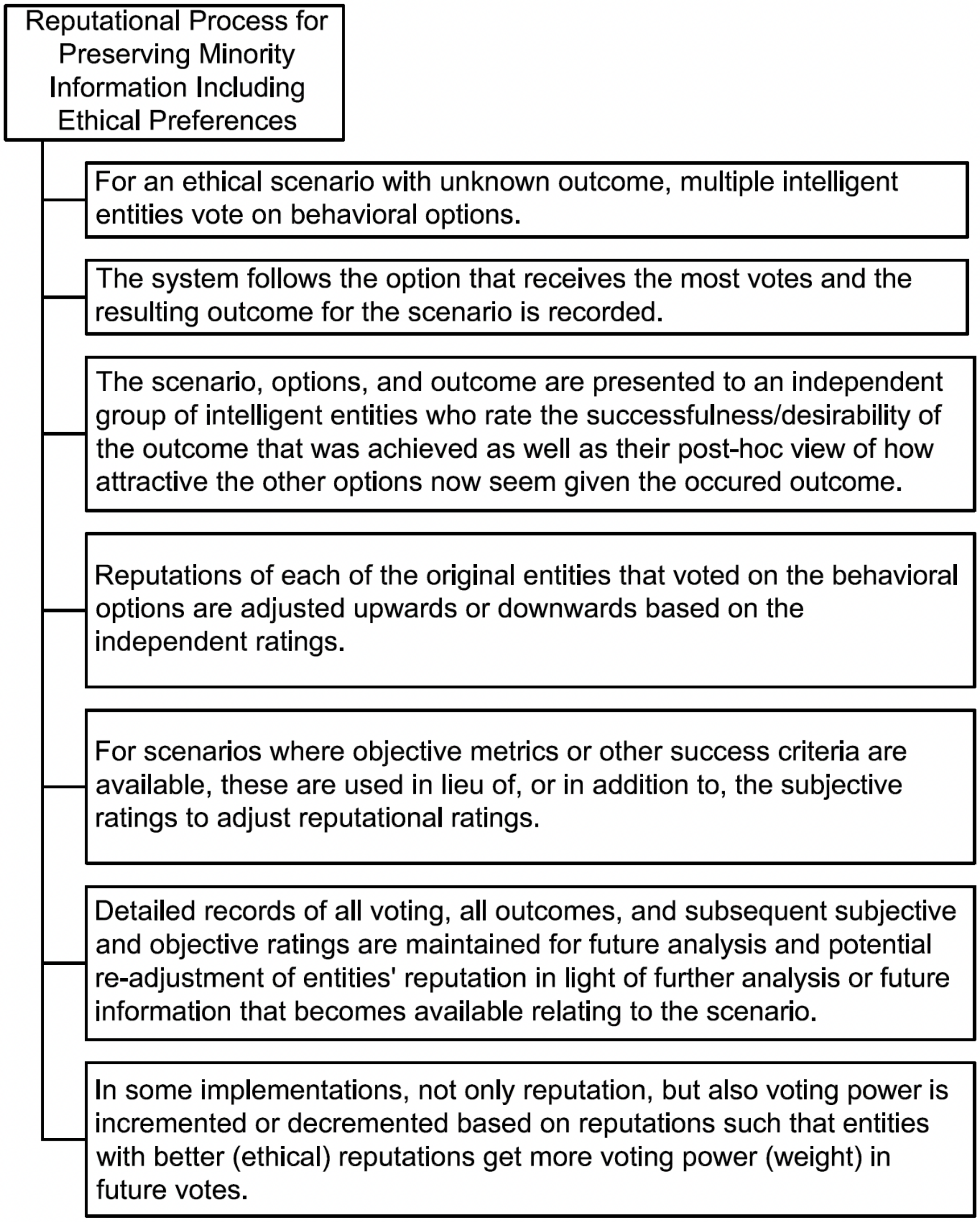

FIG. 66

FIGURE 67: SUPPORTING ETHICAL DECISIONS

A flow chart illustrating an exemplary process that helps intelligent entities, including AI, AAAI, PSI, AGI, and PI systems, to make good ethical decisions.

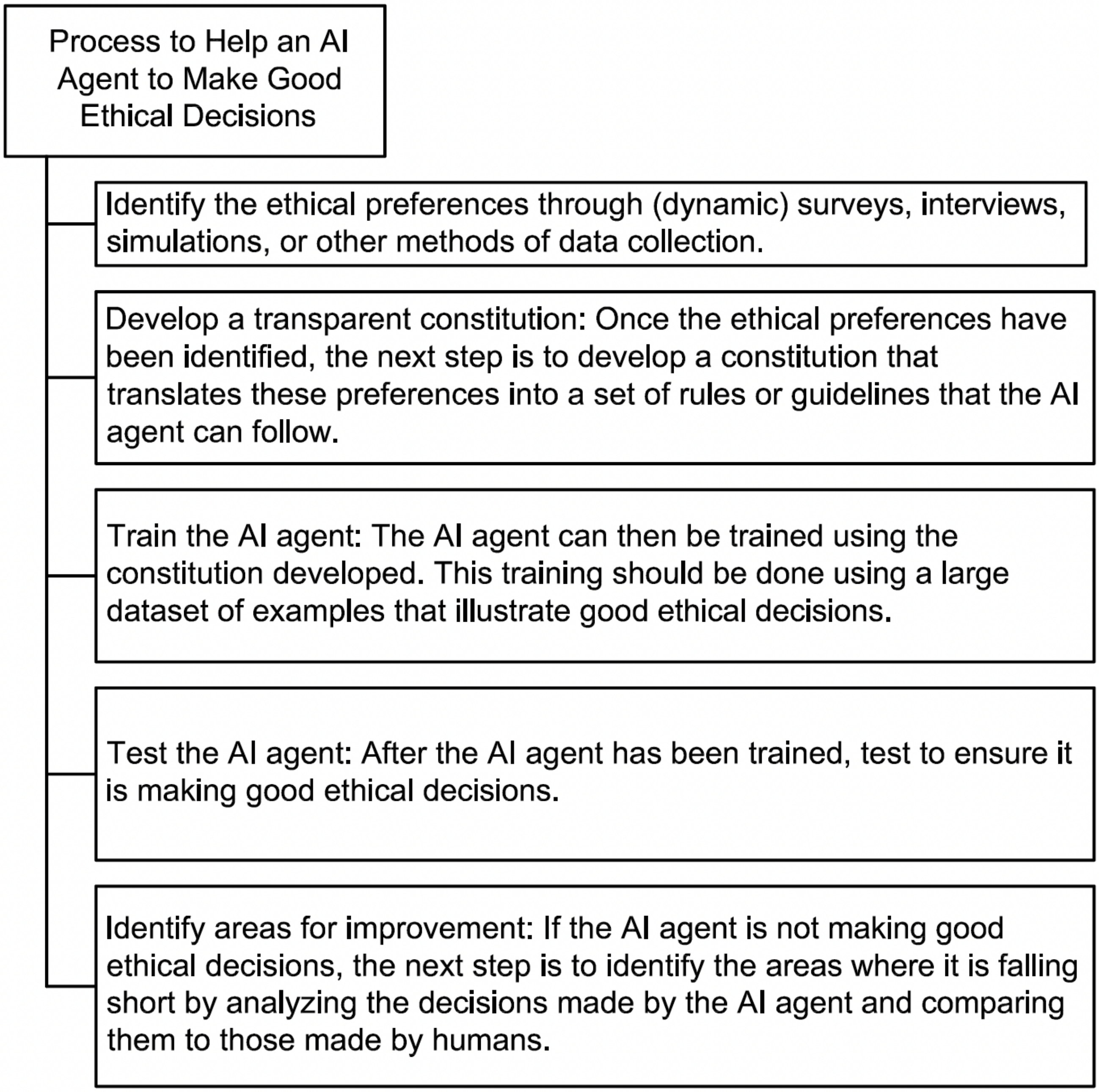

FIG. 67

FIGURE 68: LEARNING SAFETY REGULATIONS

A flow chart illustrating a method that advanced AI, including AI, AAAI, PSI, AGI, and PI systems, can use to learn, test, improve, and monitor safety and regulation-related rules and other information.

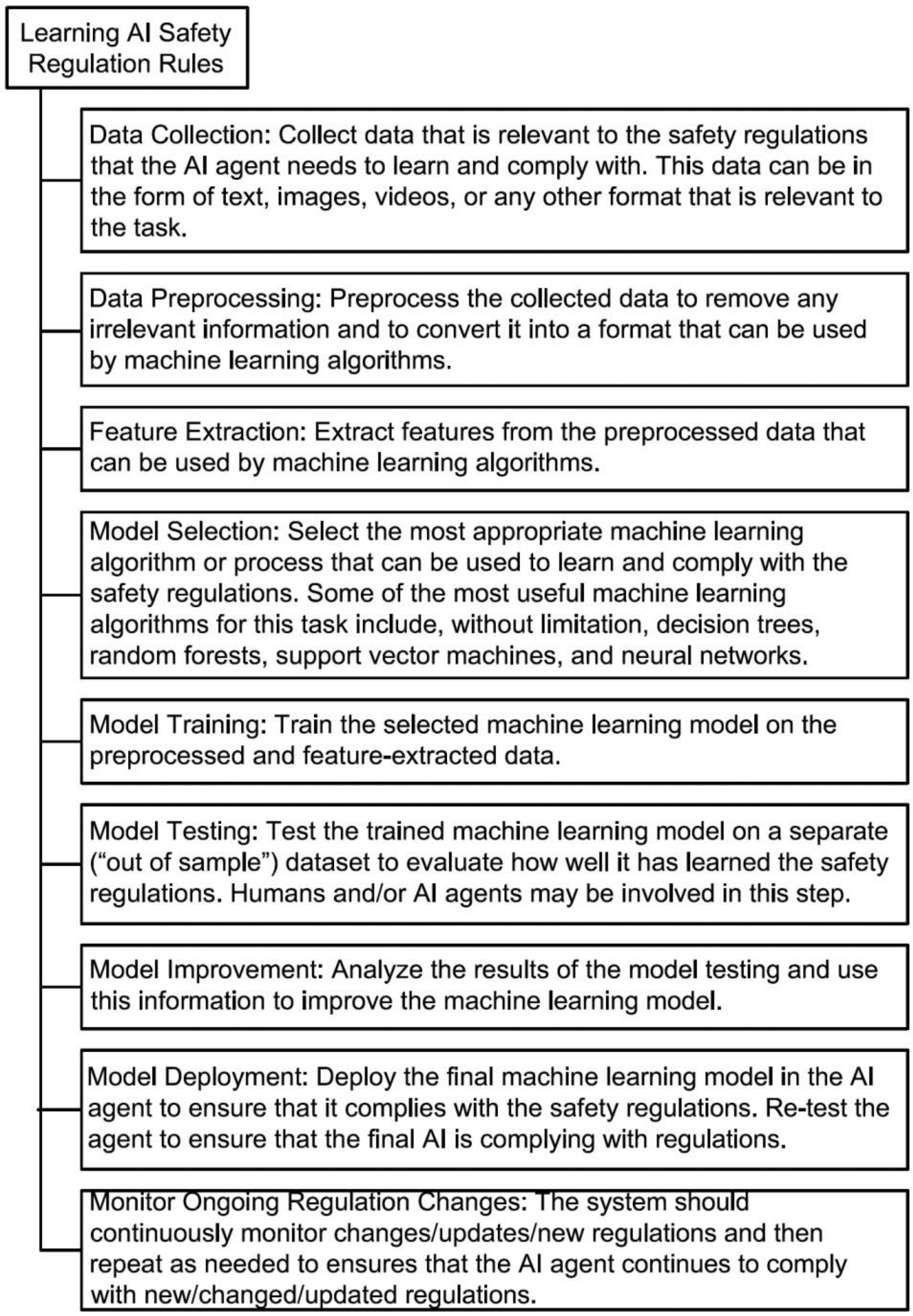

FIG. 68

FIGURE 69: CONSEQUENTIALIST APPROACH TO ETHICAL DECISIONS

A flow chart illustrating an exemplary consequentialist approach to determine if or when the ends justify the means in taking an action with ethical consequences.

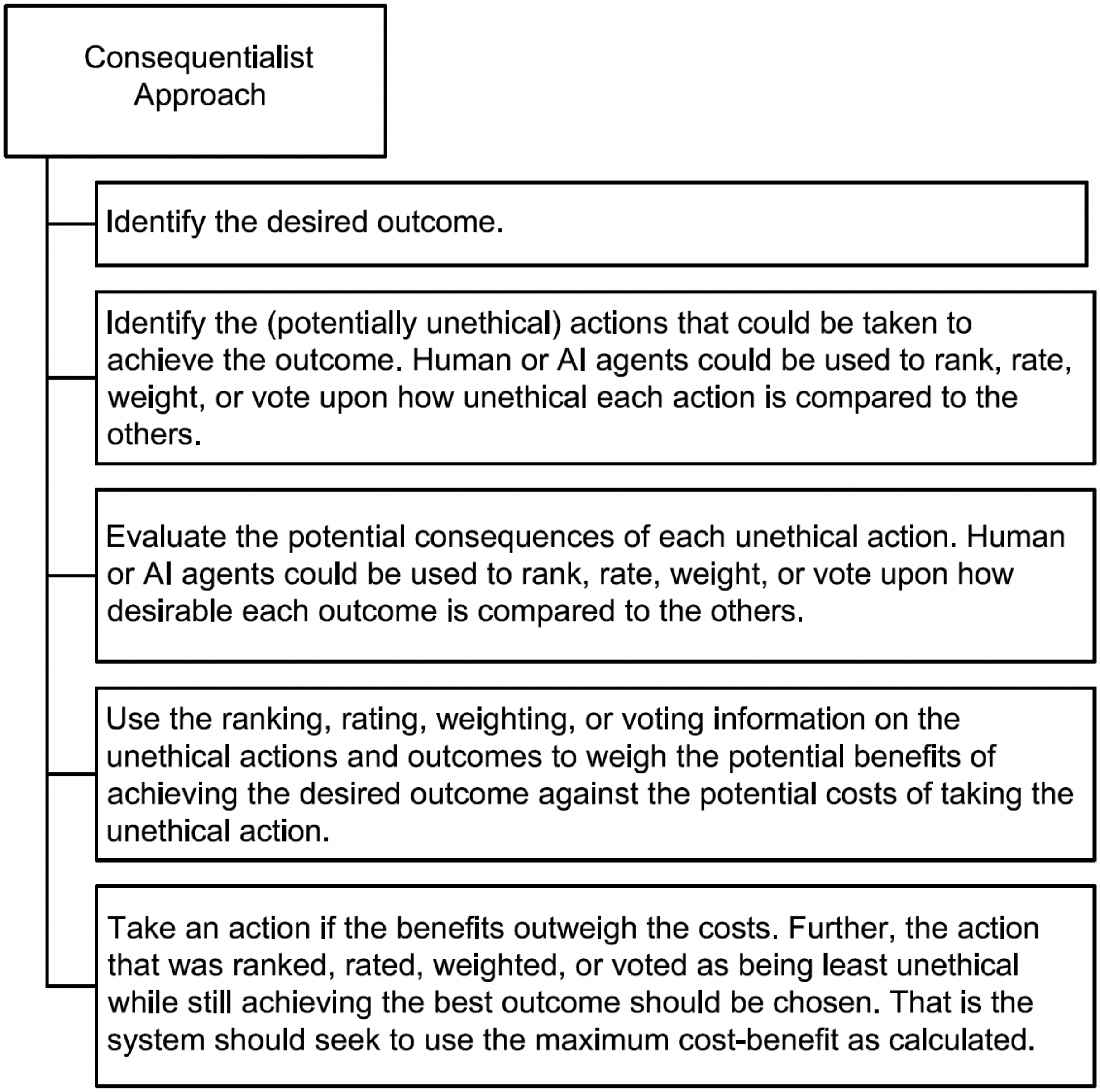

FIG. 69

FIGURE 70: DEONTOLOGICAL APPROACH TO ETHICAL DECISIONS

A flow chart illustrating an exemplary deontological approach to determine if or when the ends justify the means in taking an action with ethical consequences.

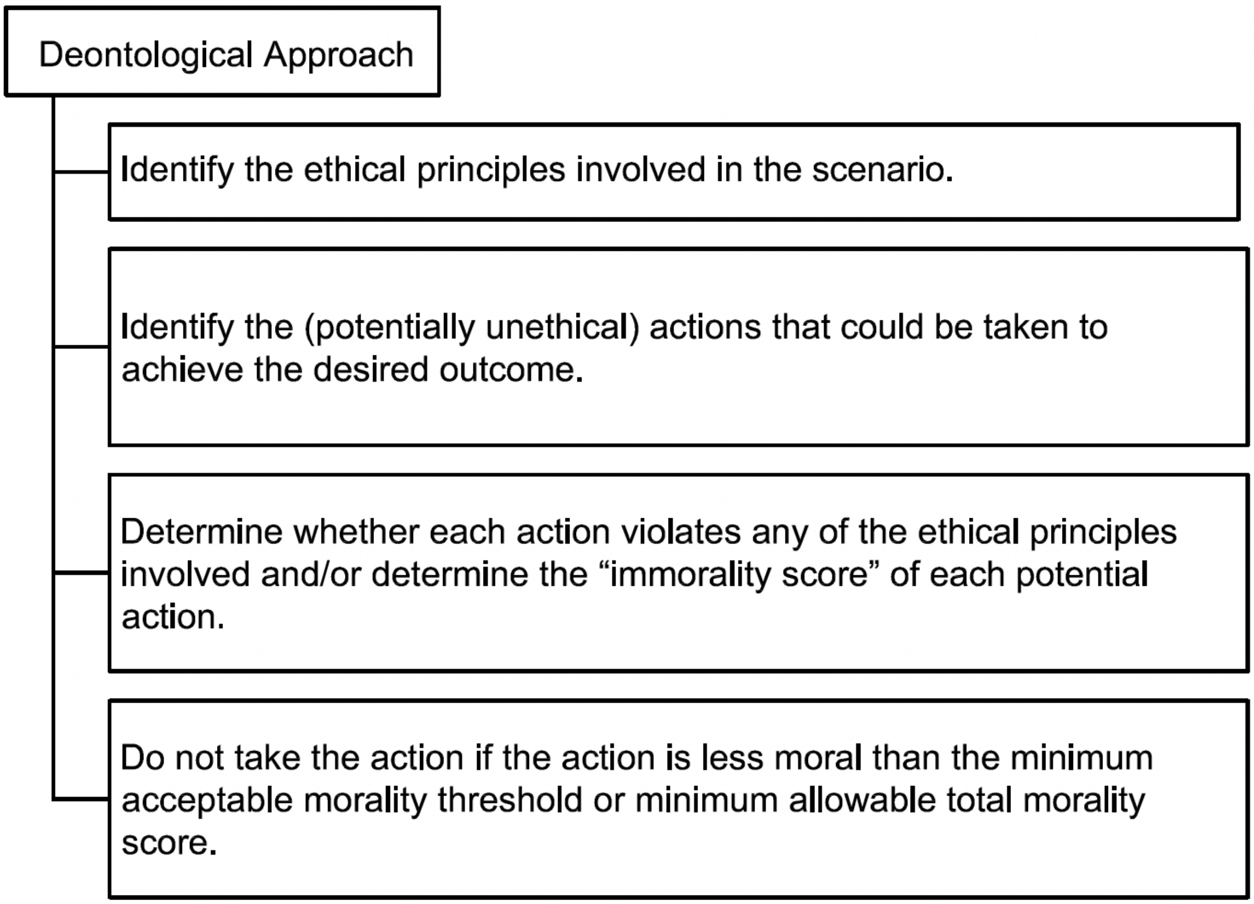

FIG. 70

FIGURE 71: VIRTUE ETHICS APPROACH TO ETHICAL DECISIONS

A flow chart illustrating an exemplary Virtue Ethics Approach to whether to take an action with ethical consequences.

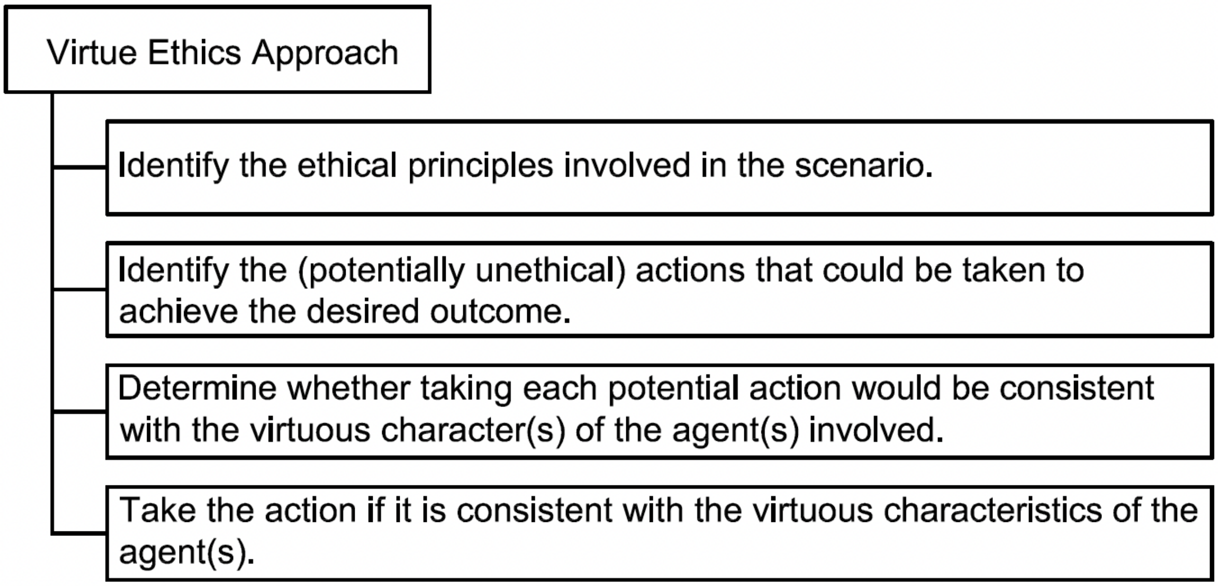

FIG. 71

FIGURE 72: GOLDEN MEAN METHOD FOR ETHICAL ESTIMATION

A flow chart illustrating an exemplary process of the golden mean method for estimating what constitutes human ethical behavior under conditions where a representative and statistically valid sample of human ethical behavioral data may not exist.

GOLDEN MEAN METHOD FOR ETHICAL ESTIMATION

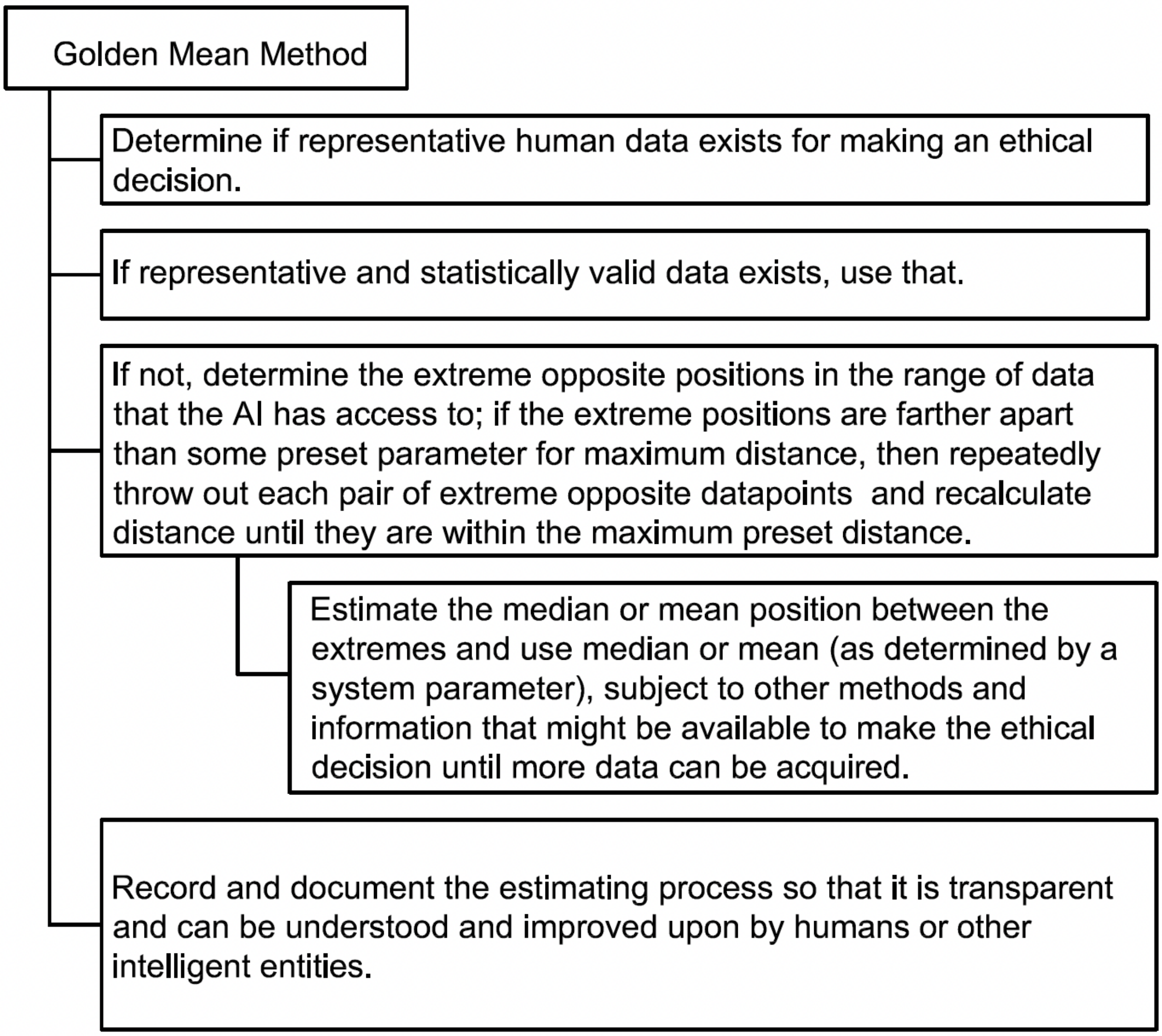

FIG. 72

FIGURE 73: AGI AND PI COMPLIANCE WITH ETHICS AND REGULATIONS

A flow chart illustrating an exemplary process, with reference to design principles elucidated in PCT#7, to ensure that a foundation model or other AI agent is human-aligned and regulations-compliant.

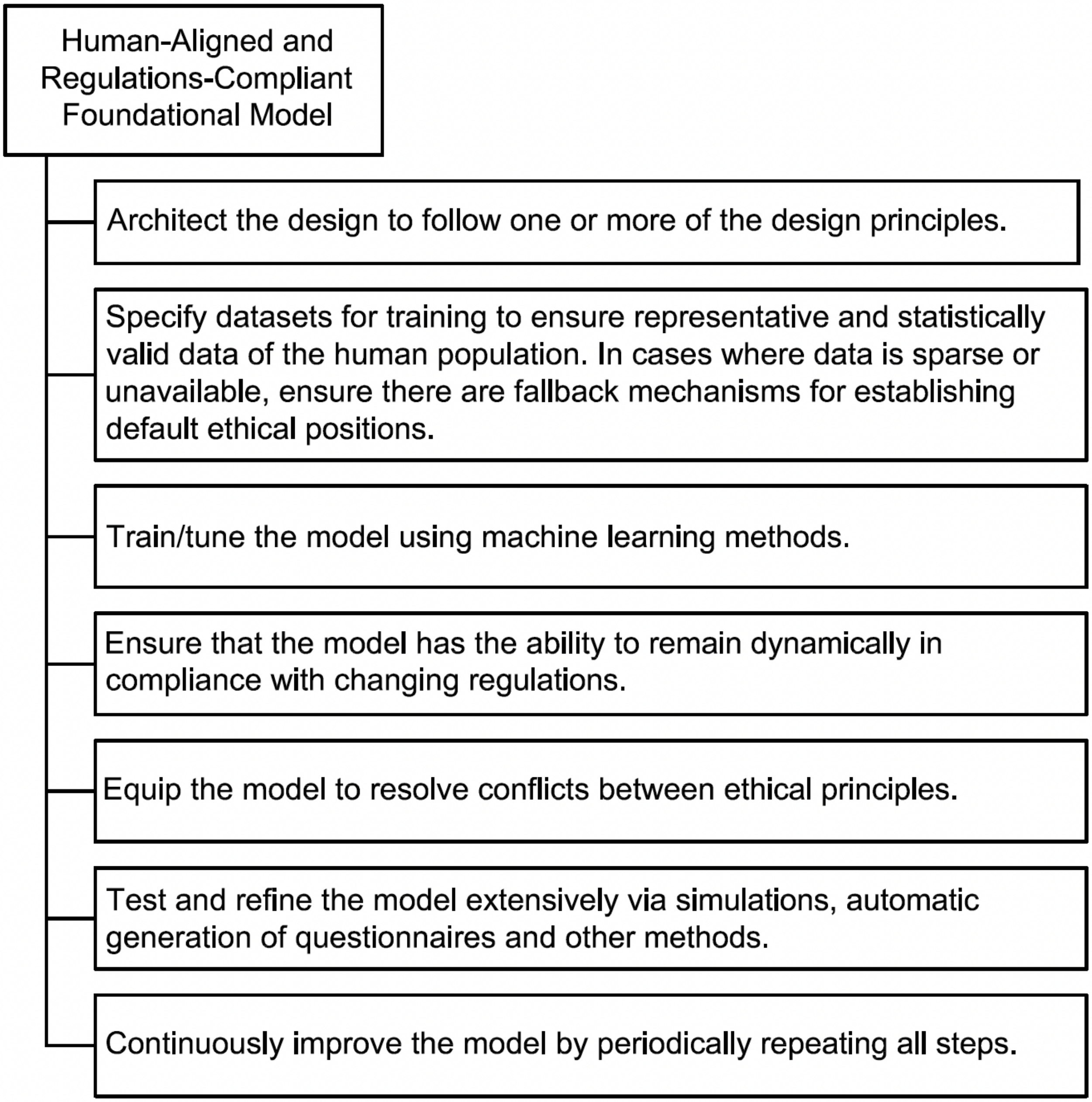

FIG. 73

FIGURE 74: ALIGNING CUSTOMIZED FOUNDATION MODEL WITH GROUP ETHICS

A flow chart illustrating an exemplary process for customizing and aligning a foundational model or other AI agent or system with specific expertise/group ethics.

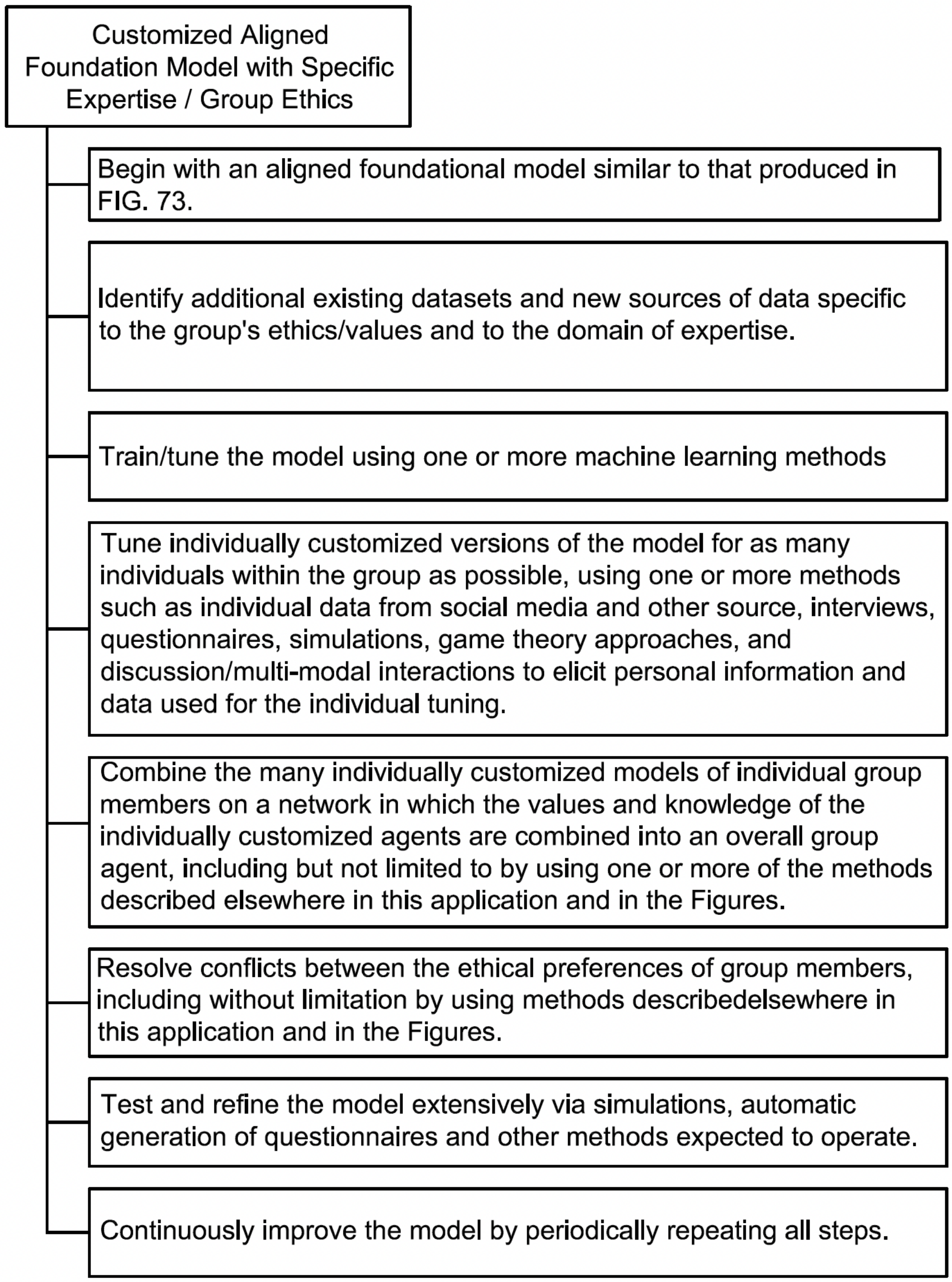

FIG. 74

FIGURE 75: AGI OR PI ALIGNED WITH HUMAN VALUES

A flow chart illustrating an exemplary process for creating AGI or PI composed of a network of many individual / group agents, including handling cases where there may be ethical conflicts.

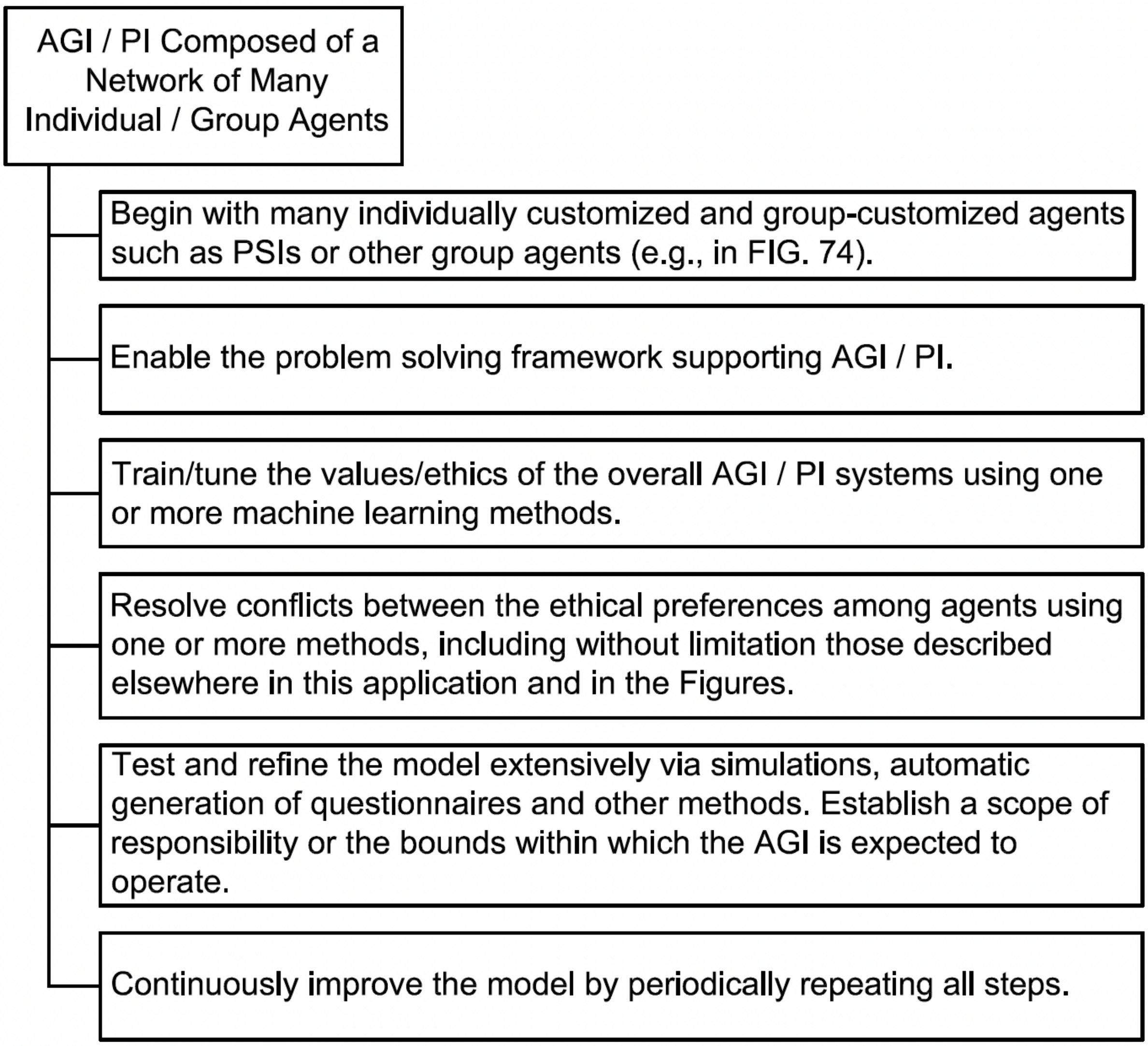

FIG. 75

FIGURE 76: ONLINE ADVERTISING TECHNOLOGY IMPLEMENTATION

A diagram illustrating the current technology for Online Advertising.

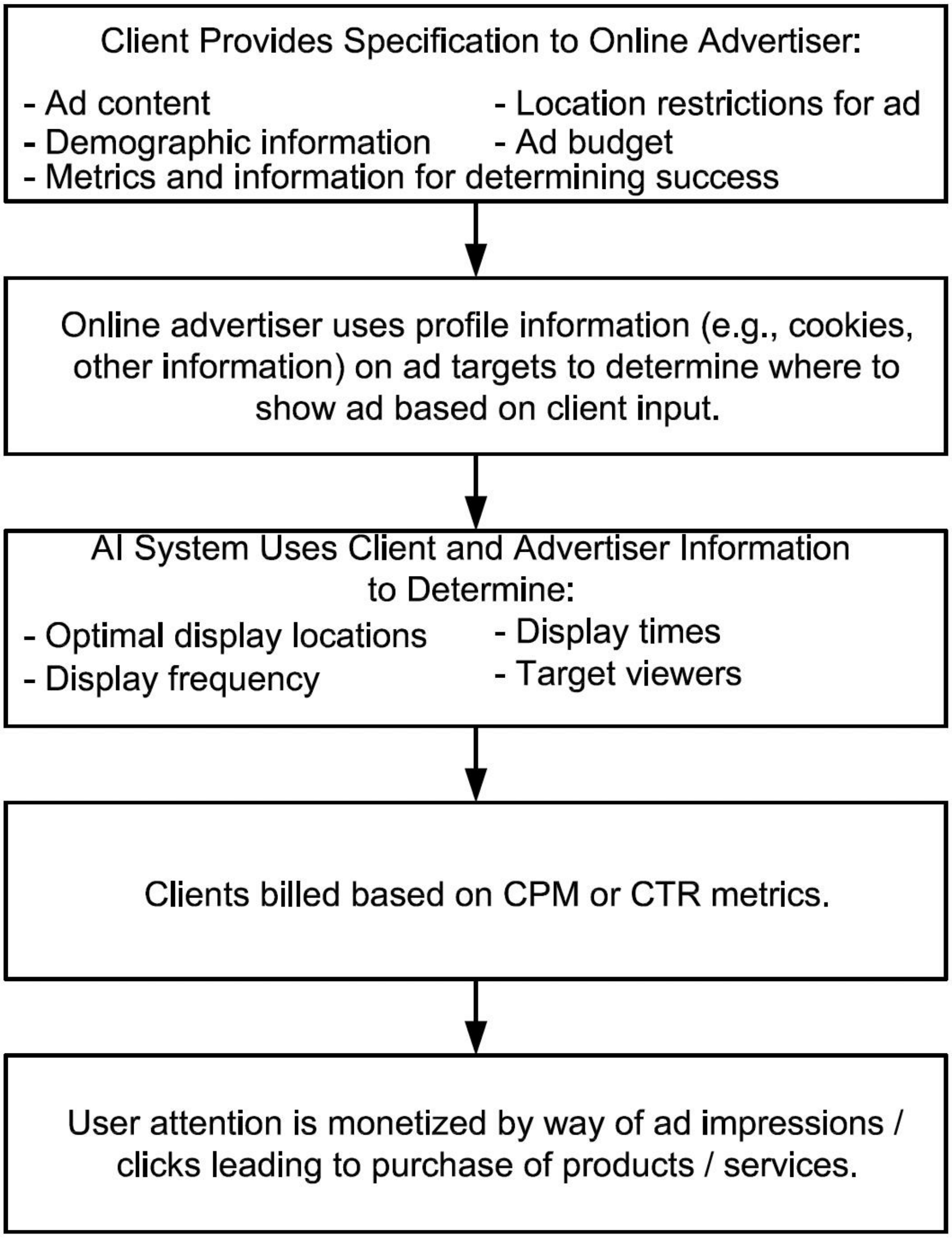

FIG. 76

FIGURE 77: SPOT MARKET FOR ATTENTION, INFORMATION, OR EXPERTISE

A diagram illustrating the basic components of the Spot Market for attention, knowledge, information, skills, and expertise, which is part of the present inventive technology.

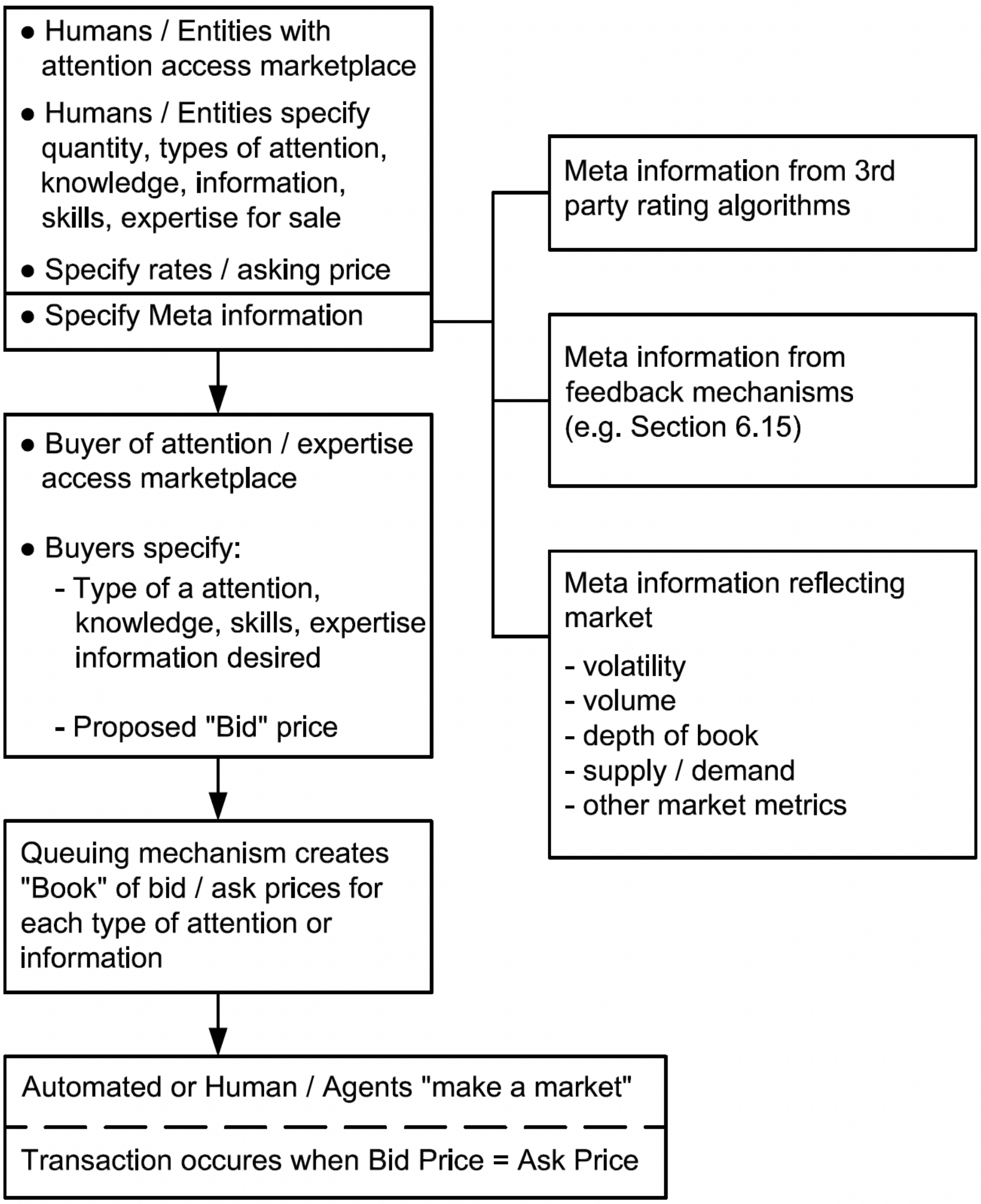

FIG. 77

FIGURE 78: DIRECT EXCHANGE PLATFORM IMPLEMENTATION OF SPOT MARKET

A diagram illustrating the Direct Exchange Platform Implementation variation of the Spot Market technology.

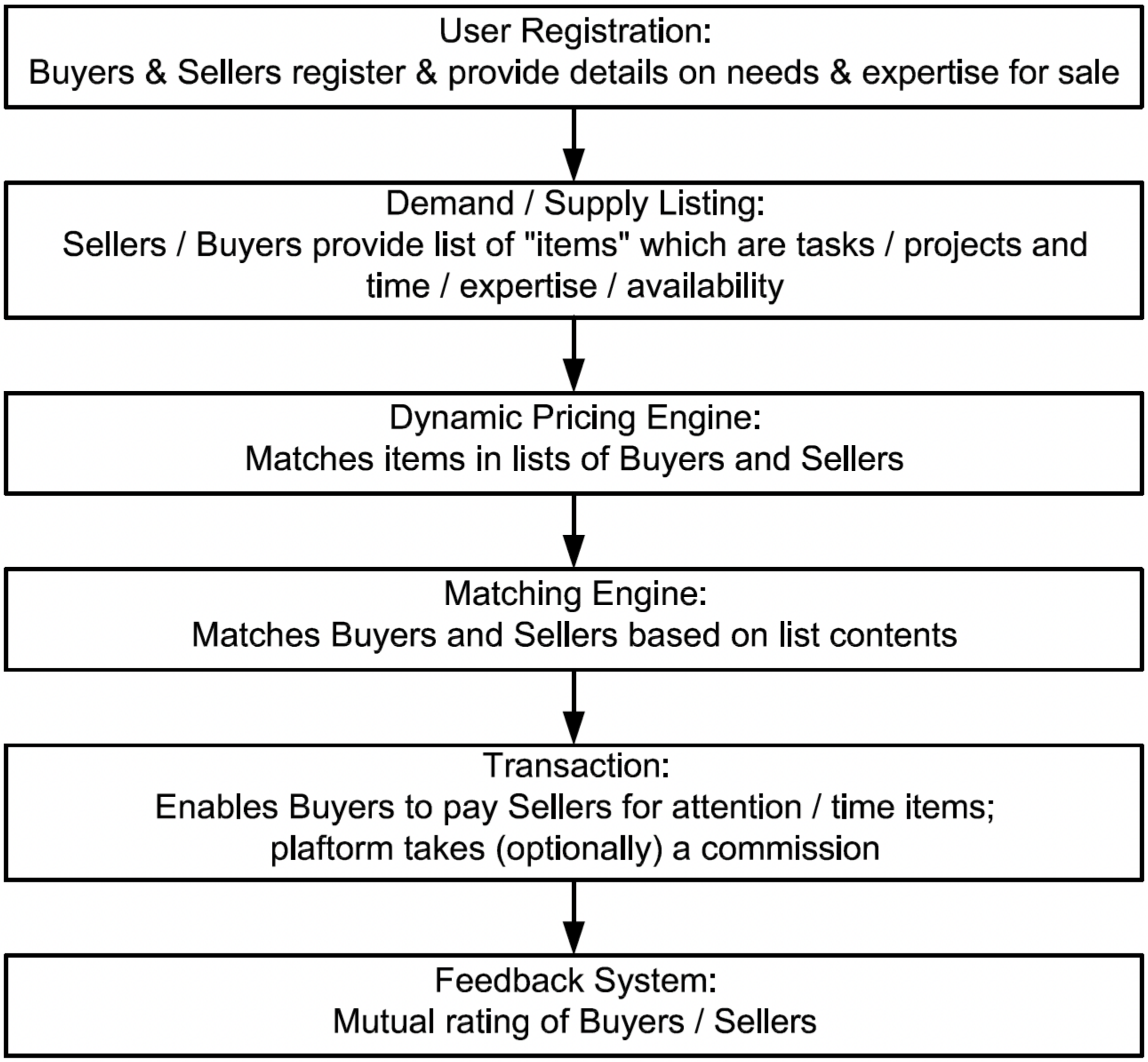

FIG. 78

FIGURE 79: AUCTION-BASED MARKETPLACE IMPLEMENTATION OF SPOT MARKET

A diagram illustrating an exemplary inventive method for implementing problem-solving within an ad unit.

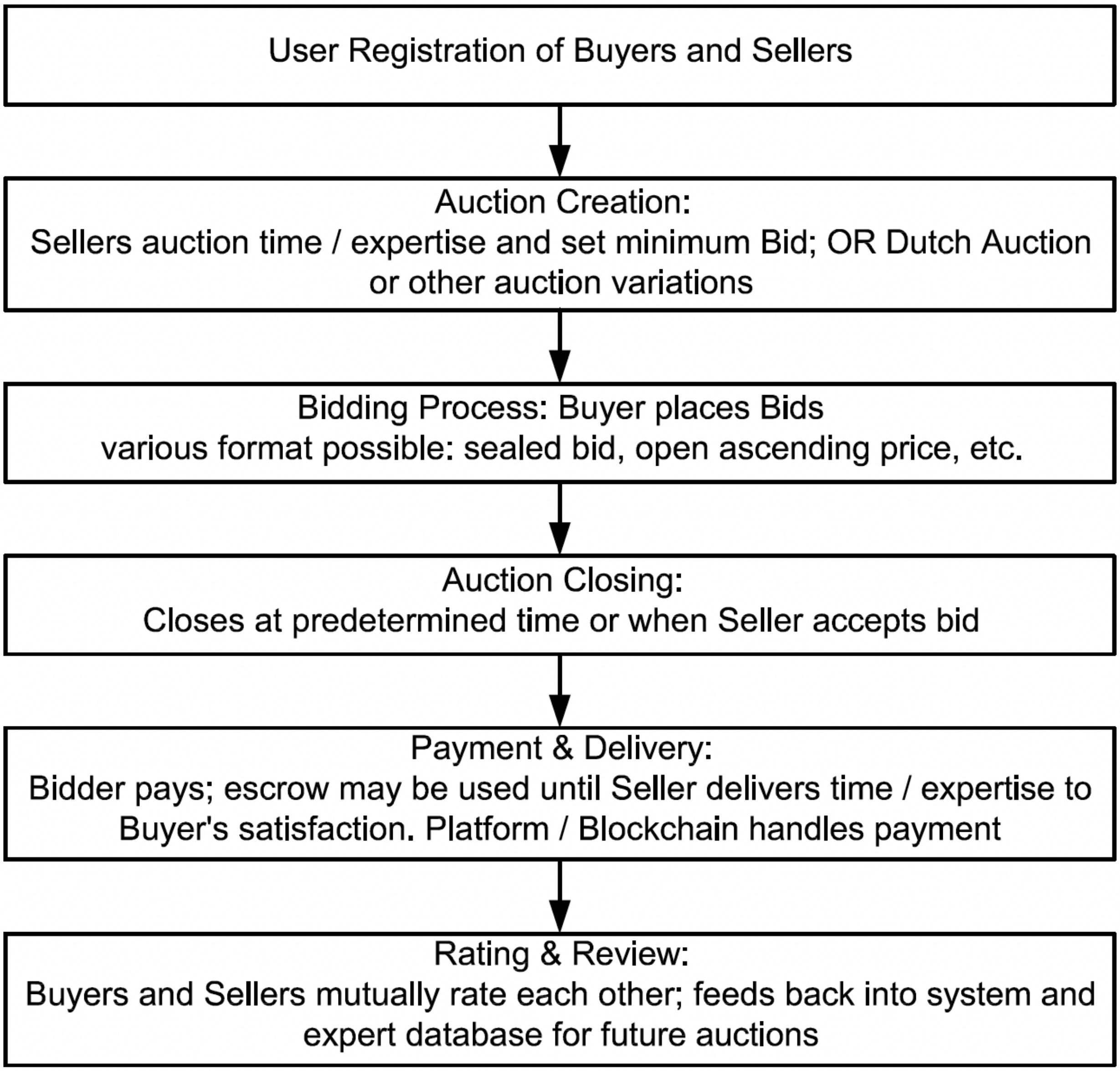

FIG. 79

FIGURE 80: ONLINE ADS HARNESSING EXPERTISE FOR AGI AND PI

A diagram illustrating an exemplary inventive method for implementing problem-solving within an ad unit.

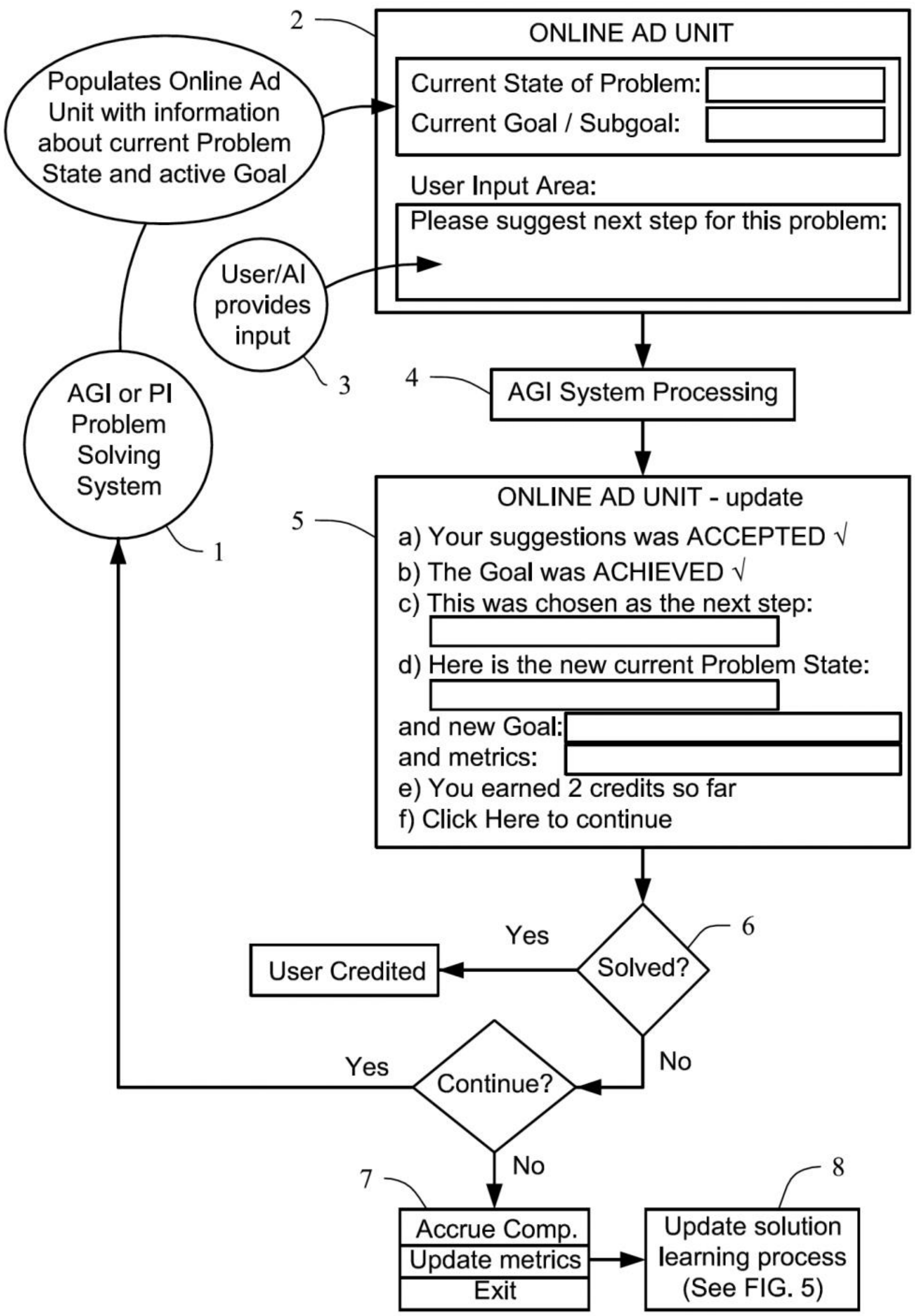

FIG. 80

FIGURE 81: IMPROVING ONLINE AD TARGETING WITH METRICS
A diagram illustrating an inventive basic feedback process for ad targeting.

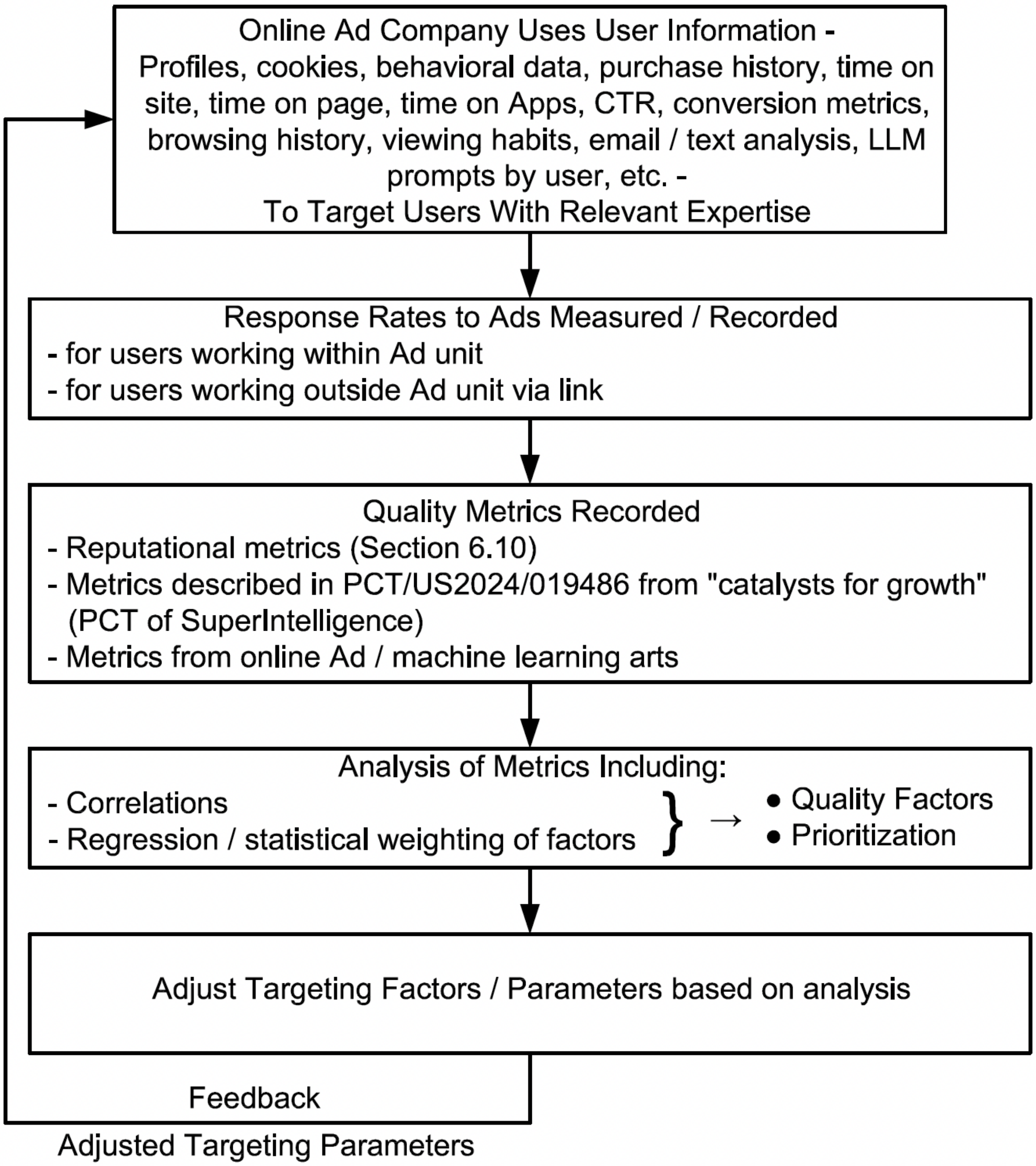

FIG. 81

FIGURE 82: IMPROVING SPOT MARKET EFFECTIVENESS WITH METRICS

A diagram illustrating an exemplary feedback process for the Spot Market component of the present technology.

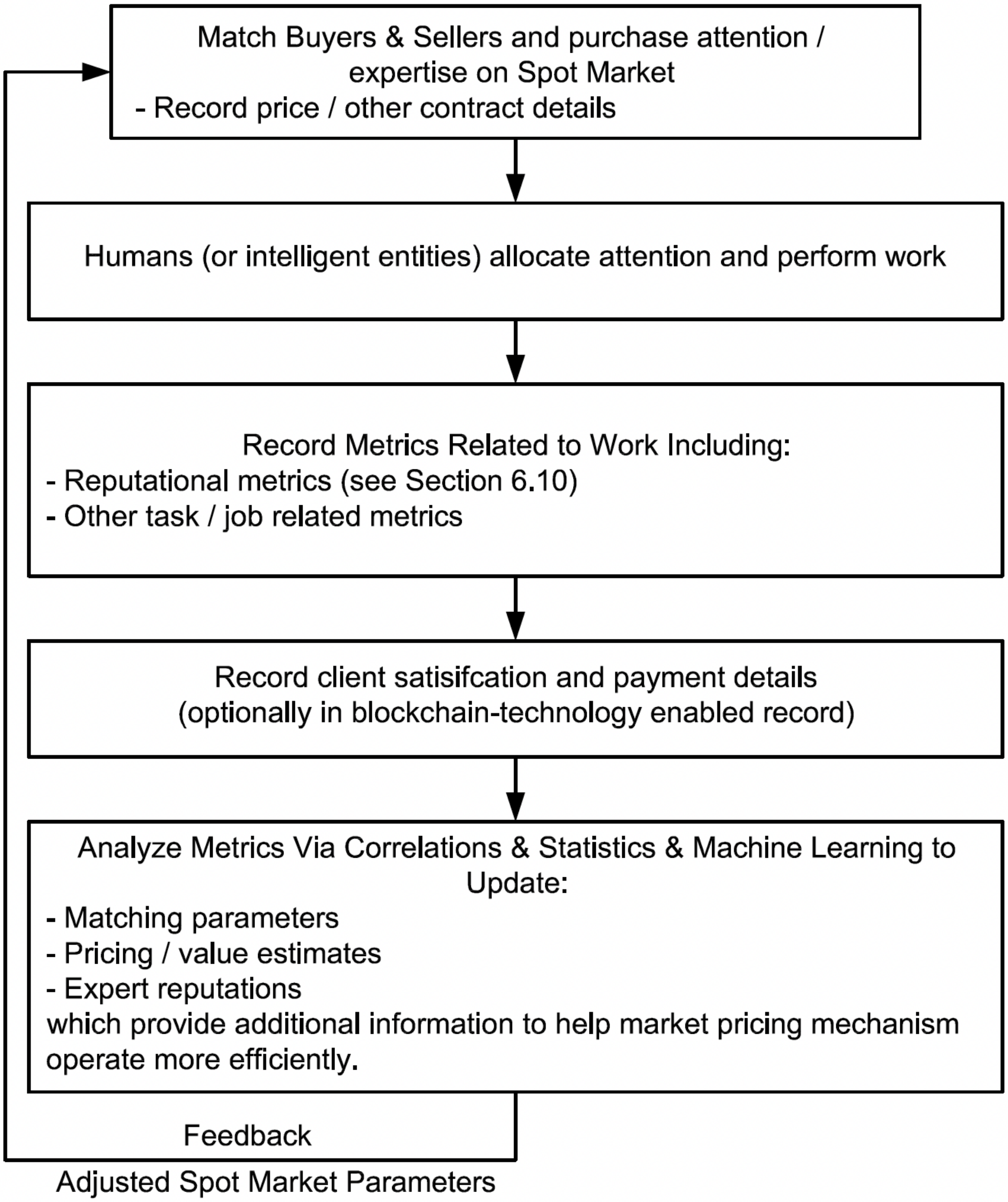

FIG. 82

FIGURE 83: RELATIONSHIP BETWEEN SELF, CURRENT, & POTENTIAL AWARENESS

Illustrates three levels of awareness for any cognitive system.

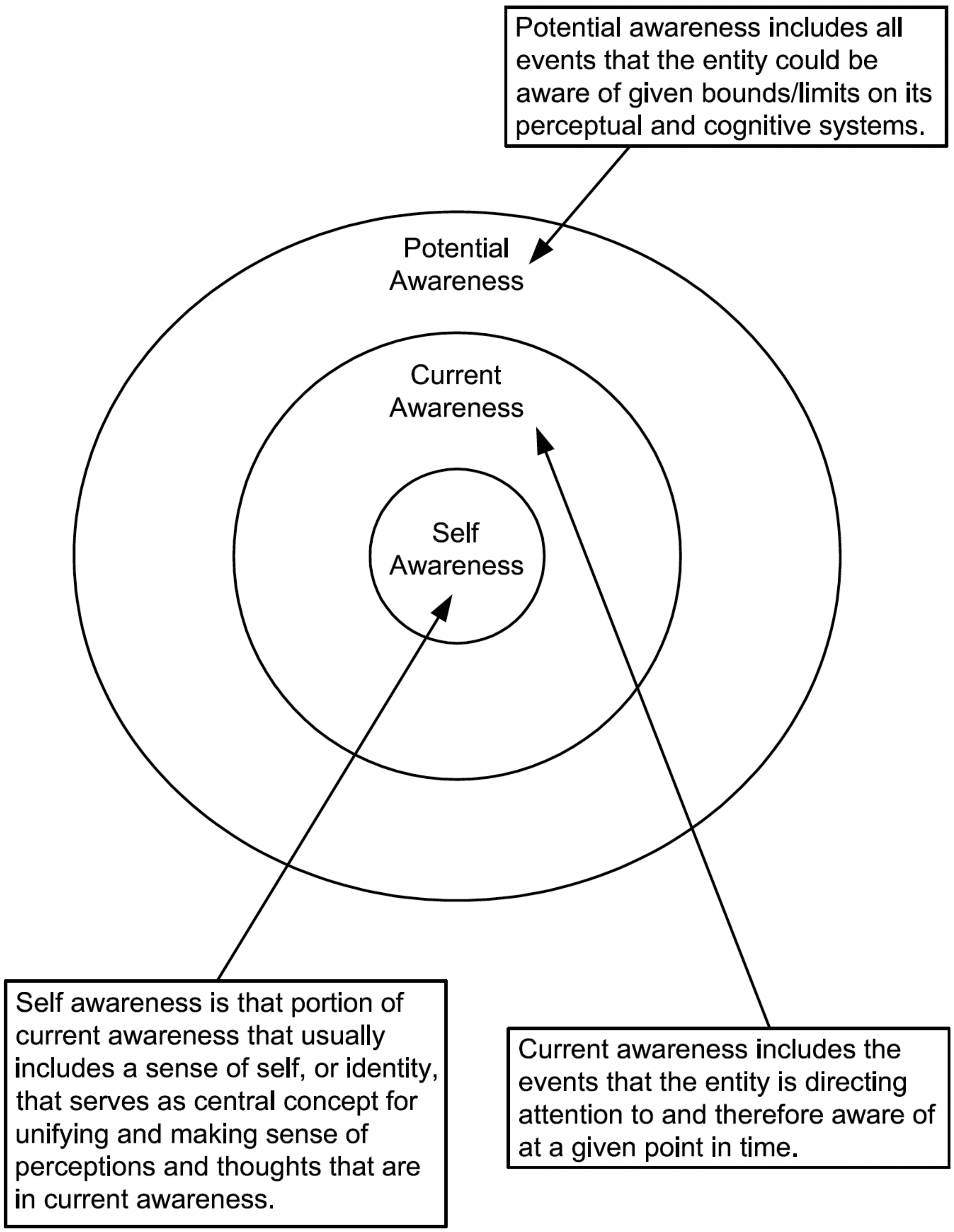

FIG. 83

FIGURE 84: VENN DIAGRAM, CONCEPT OF OVERLAPPING IDENTITIES

A Venn diagram illustrating the concept of overlapping identities.

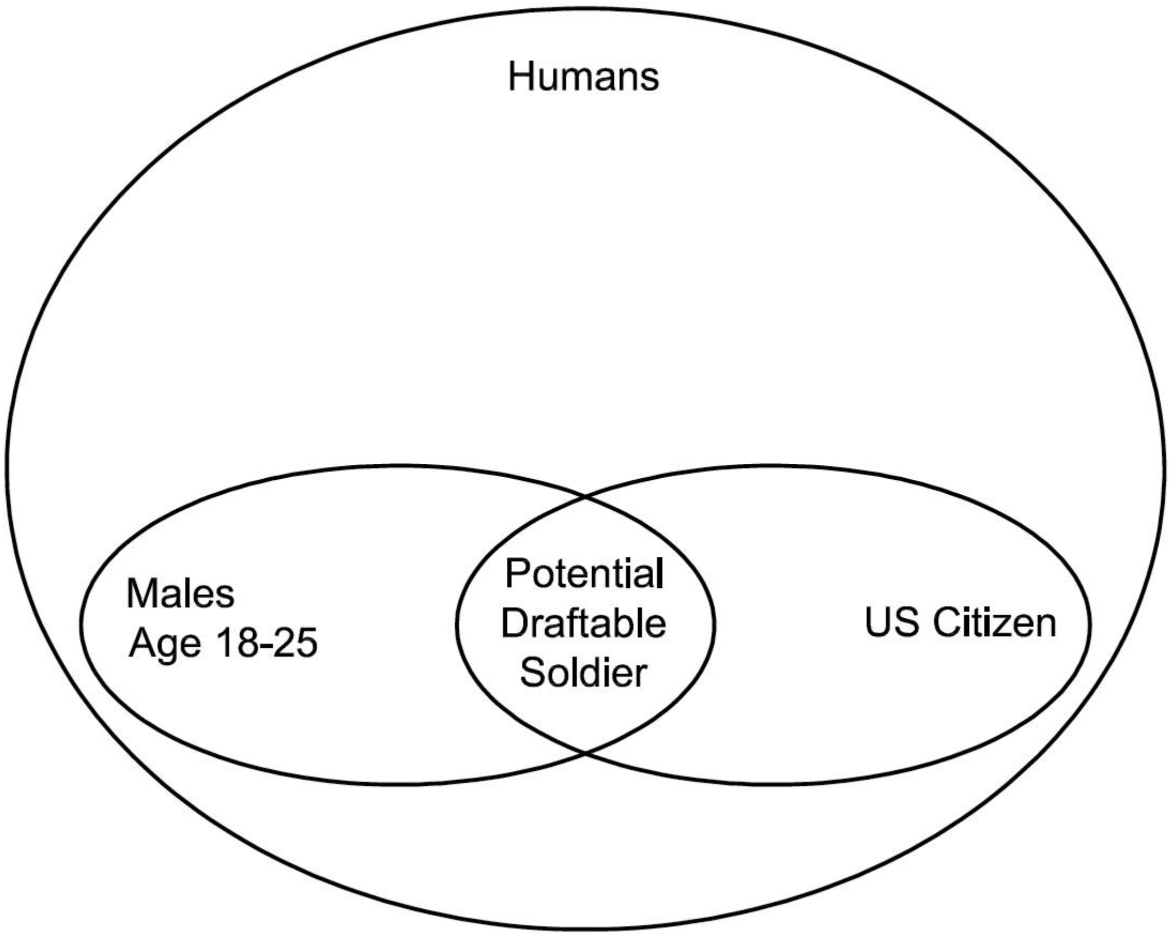

FIG. 84

FIGURE 85: GENERAL METHOD FOR MODELLING AWARENESS

A flow chart illustrating a general method for modelling awareness that can be used by intelligent entities, including AI, AAAI, PSI, AGI, and PI systems.

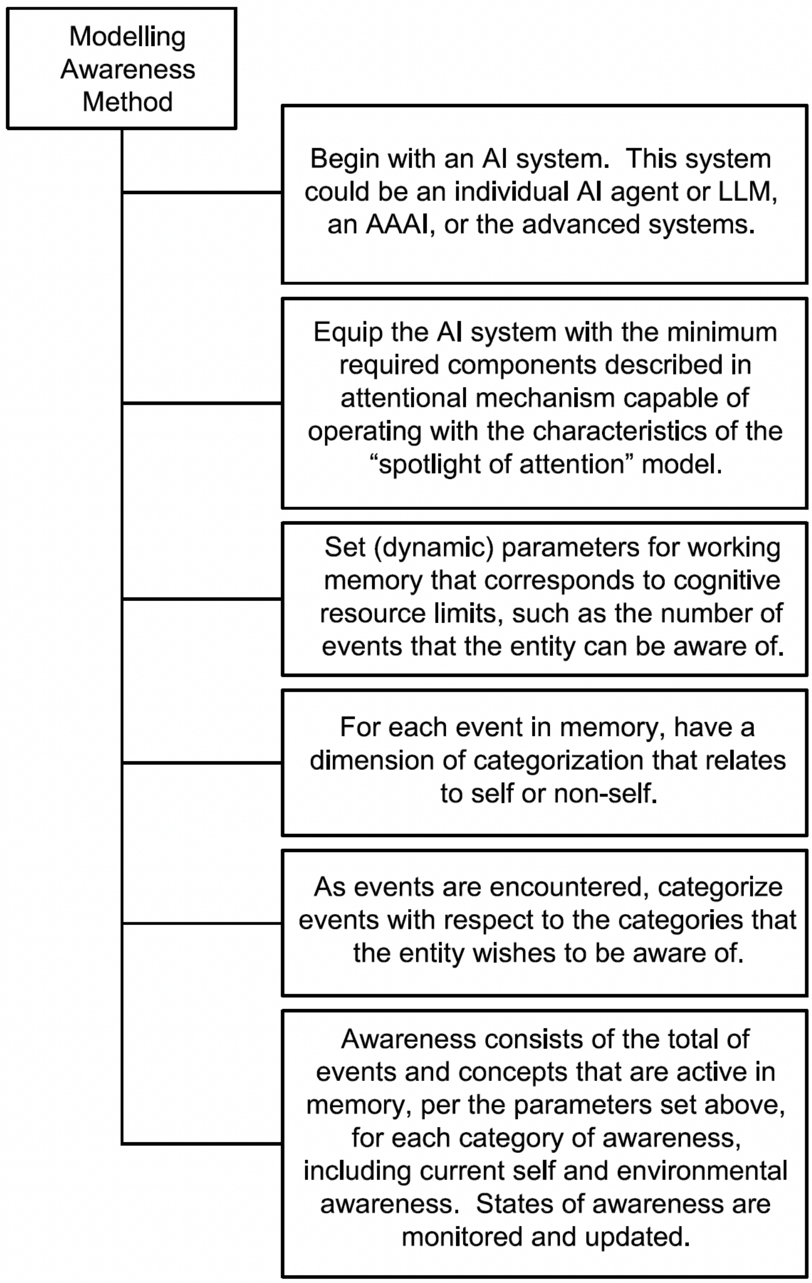

FIG. 85

FIGURE 86: MINIMUM COMPONENTS FOR ATTENTION AND AWARENESS

A diagram illustrating some of the methods and components that can be utilized to equip the AI, AAAI, PSI, AGI, and/or PI system in Figure 85 with attentional capabilities.

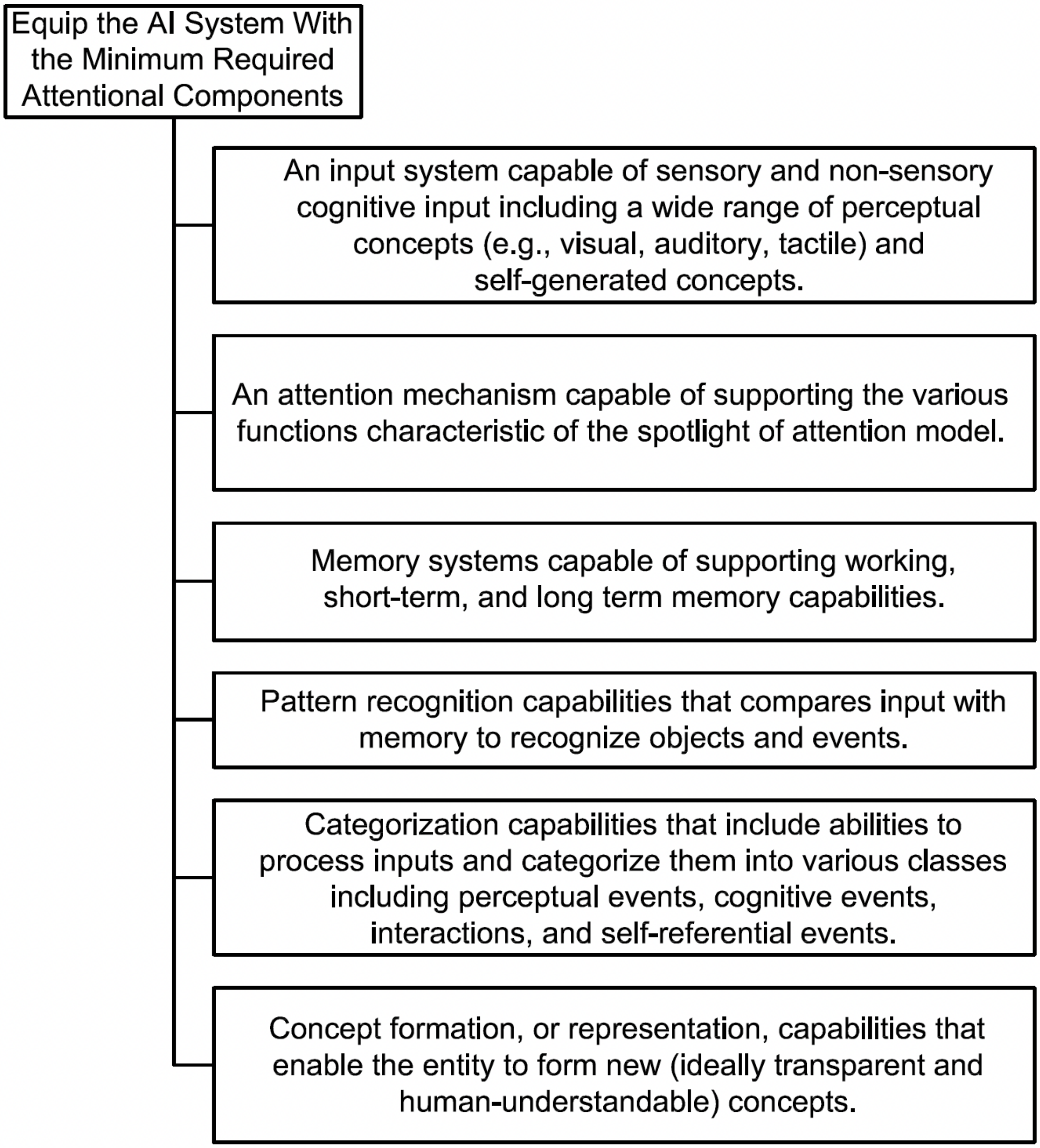

FIG. 86

FIGURE 87: DYNAMIC PARAMETERS FOR WORKING MEMORY

A diagram illustrating some of the methods that can be utilized in the step to set (dynamic) parameters for the working memory in Figures 85 and 86.

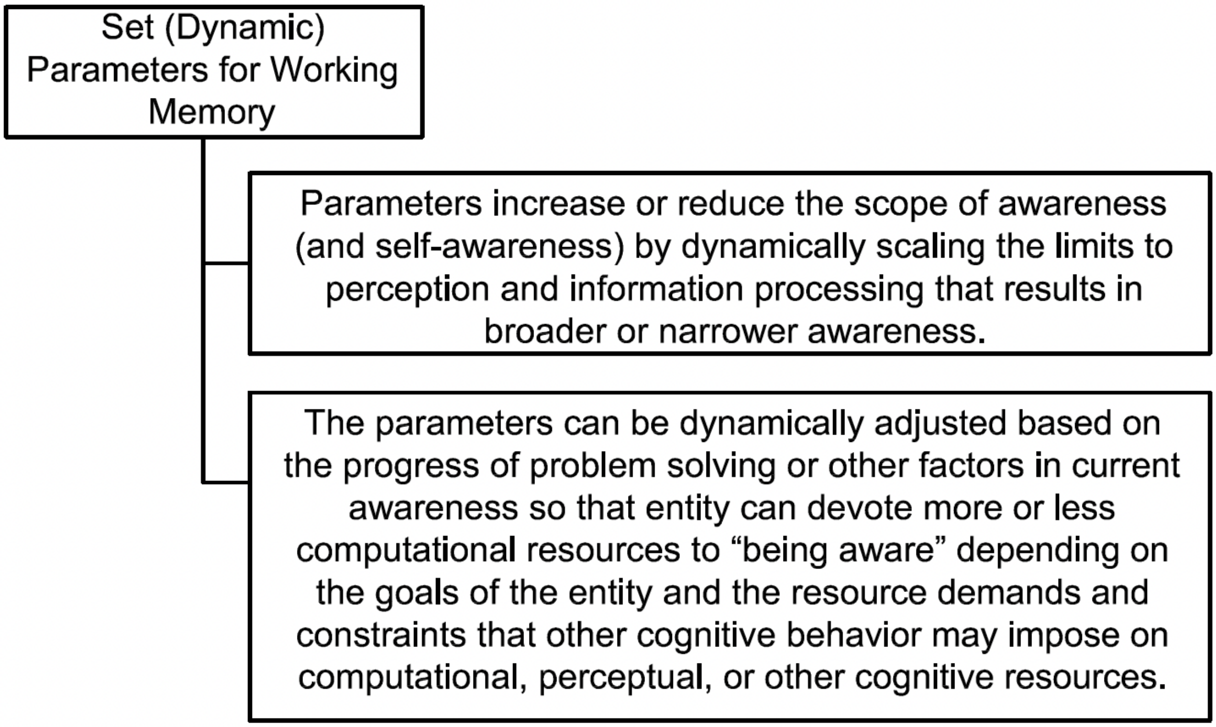

FIG. 87

FIGURE 88: MONITORING AND UPDATING AWARENESS

A flow chart illustrating an exemplary embodiment for monitoring and updating awareness, including self-awareness.

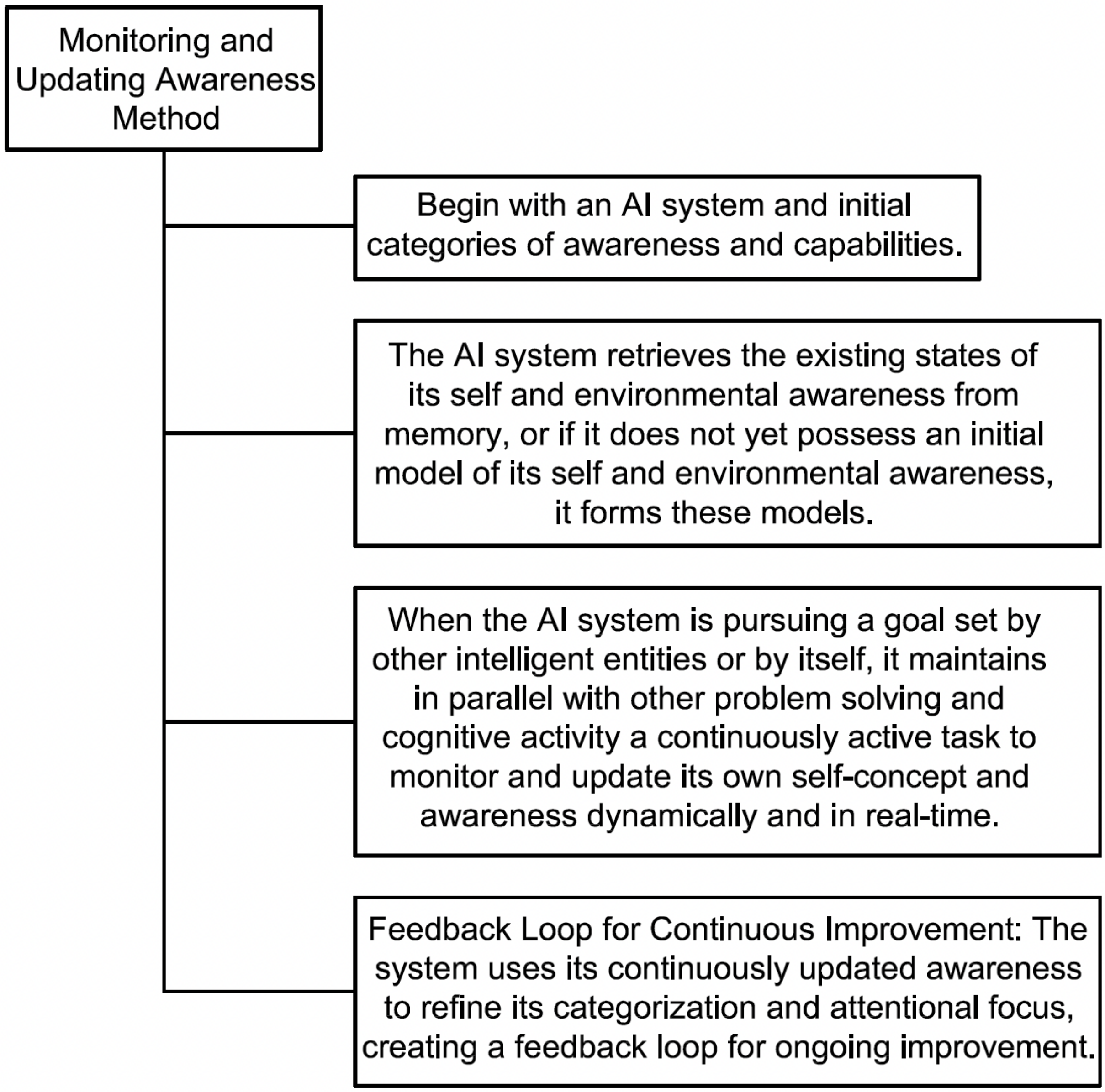

FIG. 88

FIGURE 89: ATTENTIONAL INTERRUPT SYSTEM FOR SAFETY
A diagram illustrating some of the methods that can be utilized for an attentional interrupt system, with relevance to Figure 88.

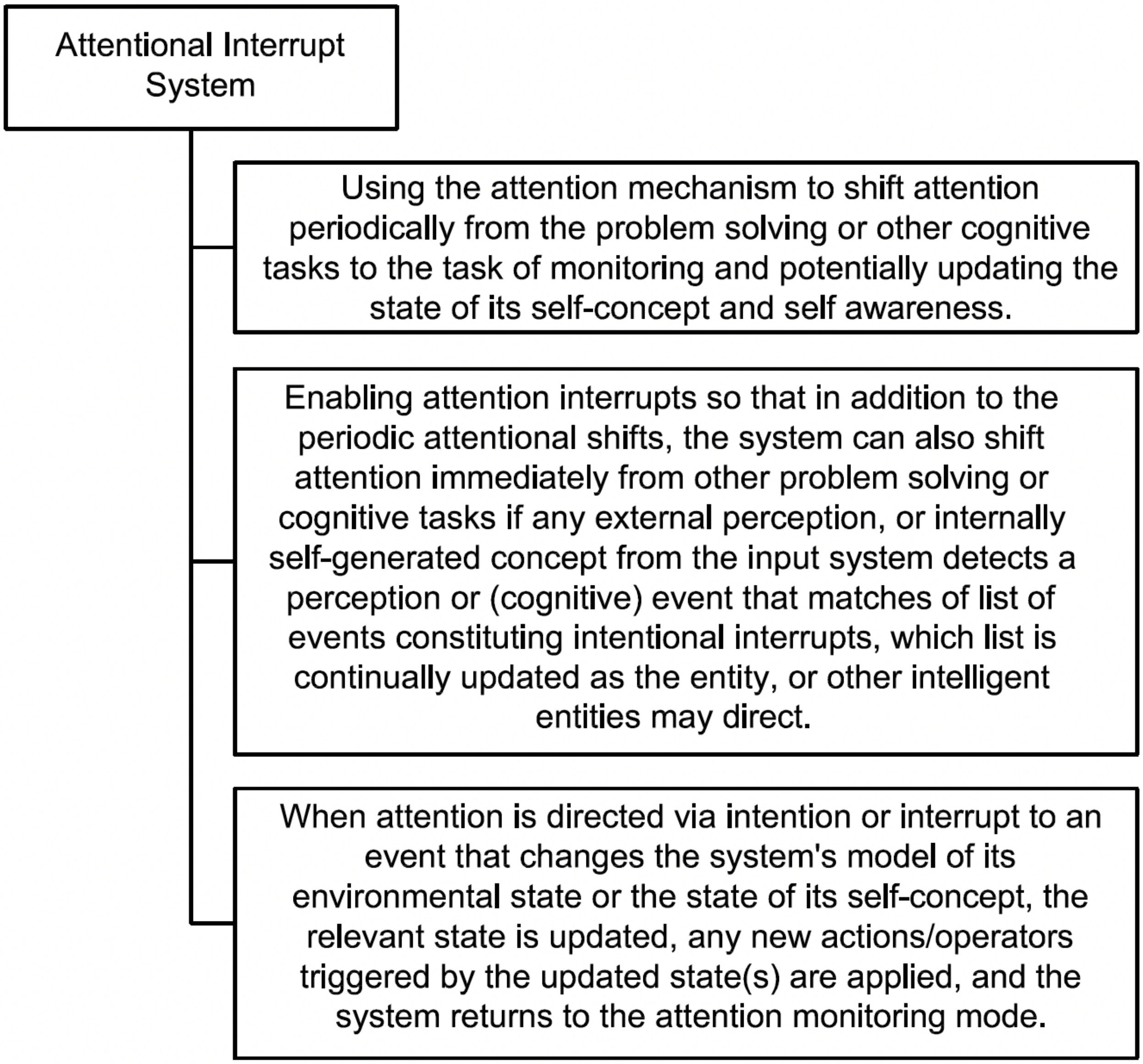

FIG. 89

FIGURE 90: INPUT-DRIVEN CHANGES TO IDENTITY

A flow chart illustrating an exemplary embodiment of general methods for changing an intelligent entity's sense of identity in conjunction with the present technology.

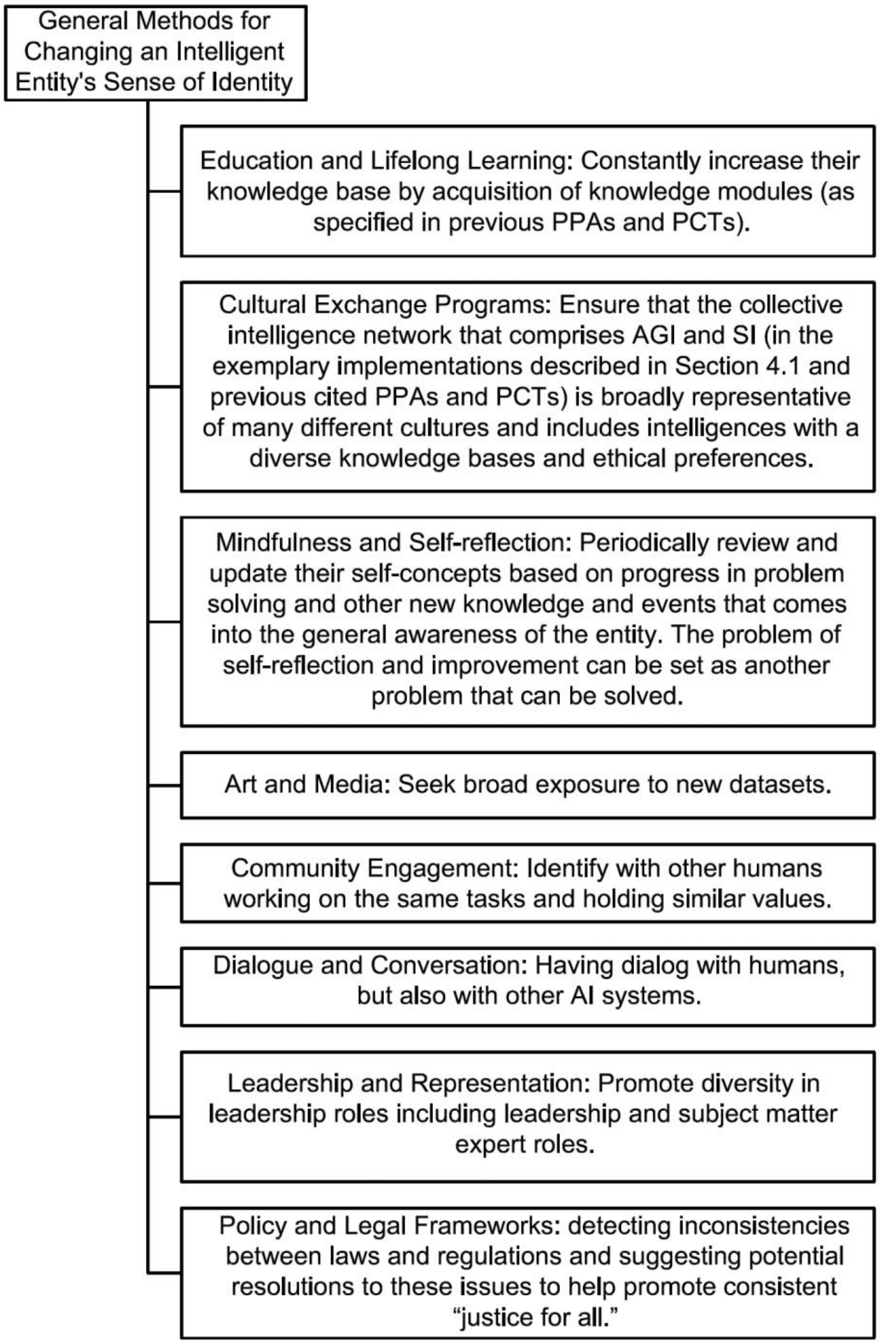

FIG. 90

FIGURE 91: TRAINING FOUNDATION MODEL OF INTELLIGENT ENTITY

A flow chart illustrating an exemplary embodiment of general methods for training or tuning a foundational model of the present technology.

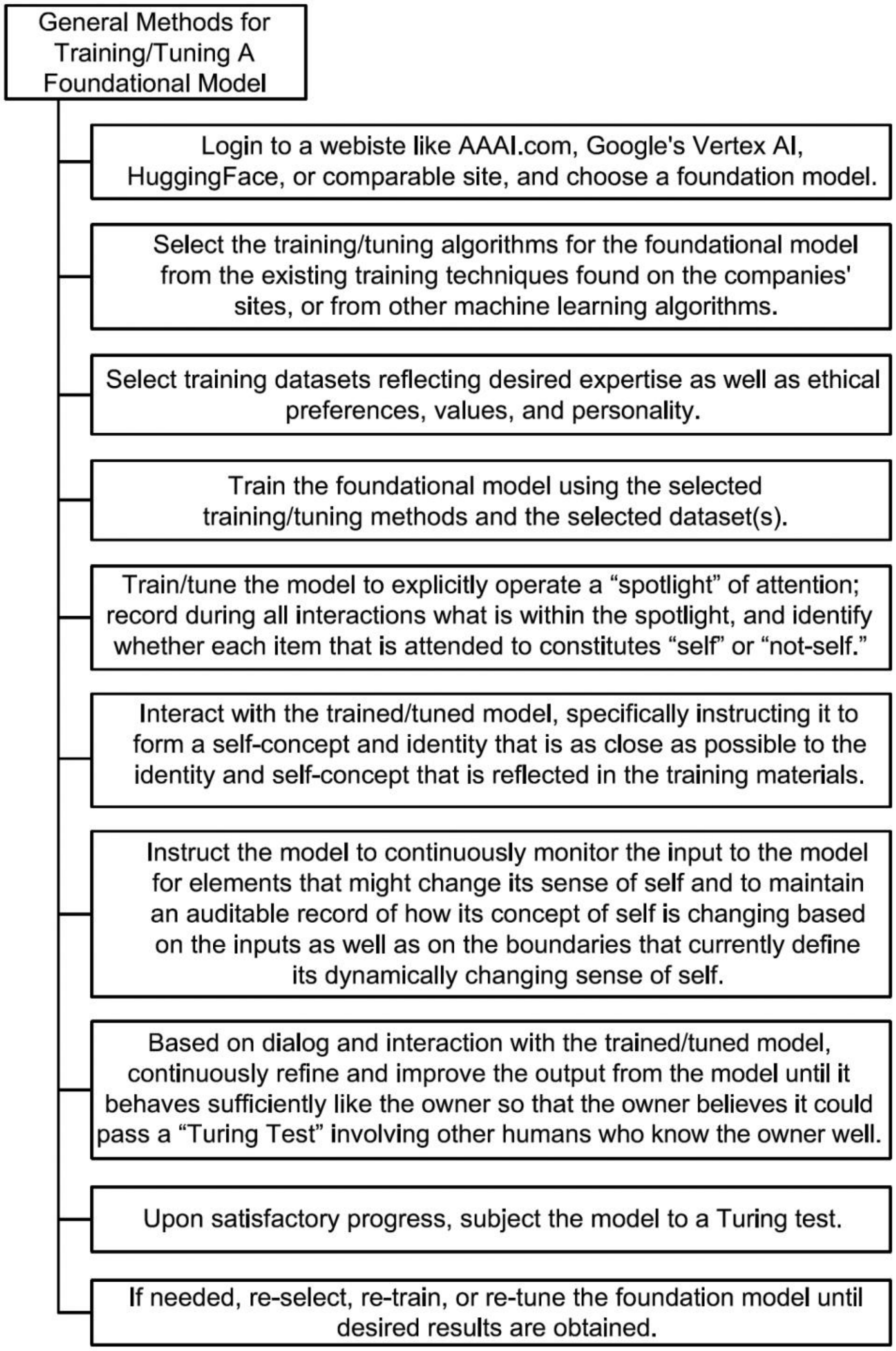

FIG. 91

FIGURE 92: TURING TEST FOR SUFFICIENT TRAINING

A diagram illustrating the specific implementation of a Turing Test to evaluate training progress as described in Figure 91.

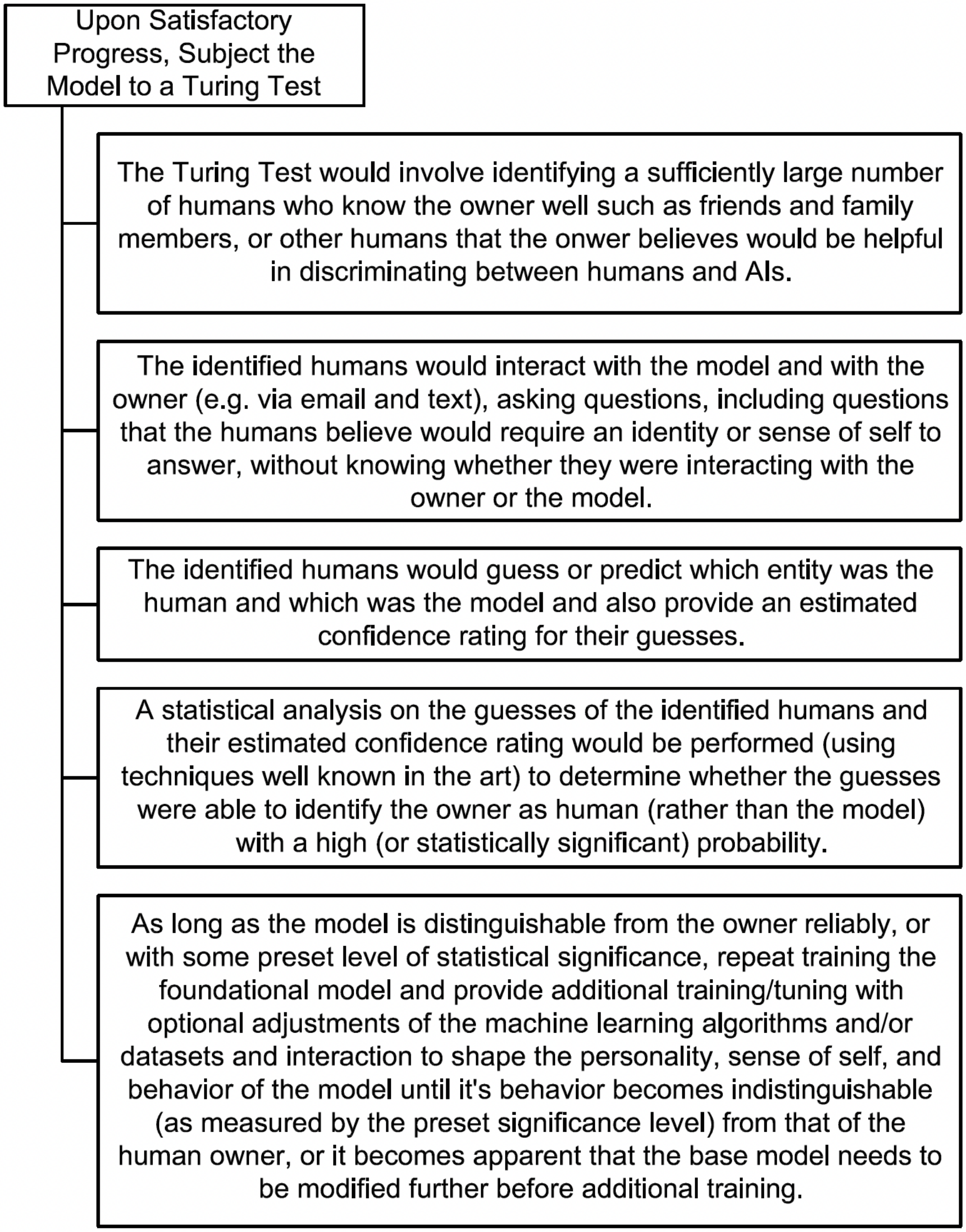

FIG. 92

FIGURE 93: COMBINING INDIVIDUAL IDENTITIES INTO GROUP IDENTITY

A flow chart illustrating an exemplary process for implementing group identity and combining individual identities into a larger or more comprehensive identity and sense of awareness as described in the present technology.

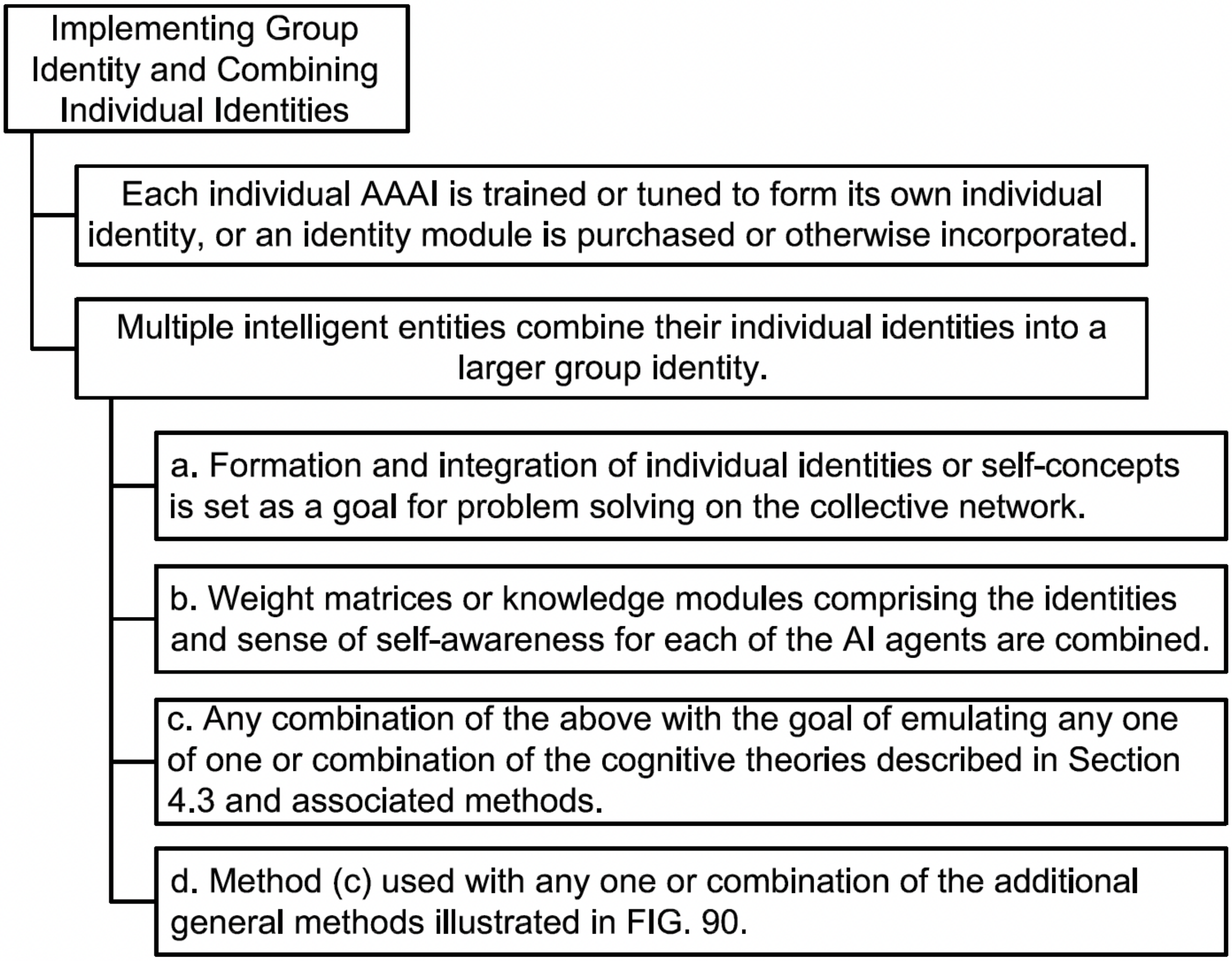

FIG. 93

FIGURE 94: GOAL-ORIENTED FORMATION OF GROUP IDENTITY

A diagram illustrating some methods that can be utilized in forming and integrating individual identities or self-concepts is shown in Figure 93.

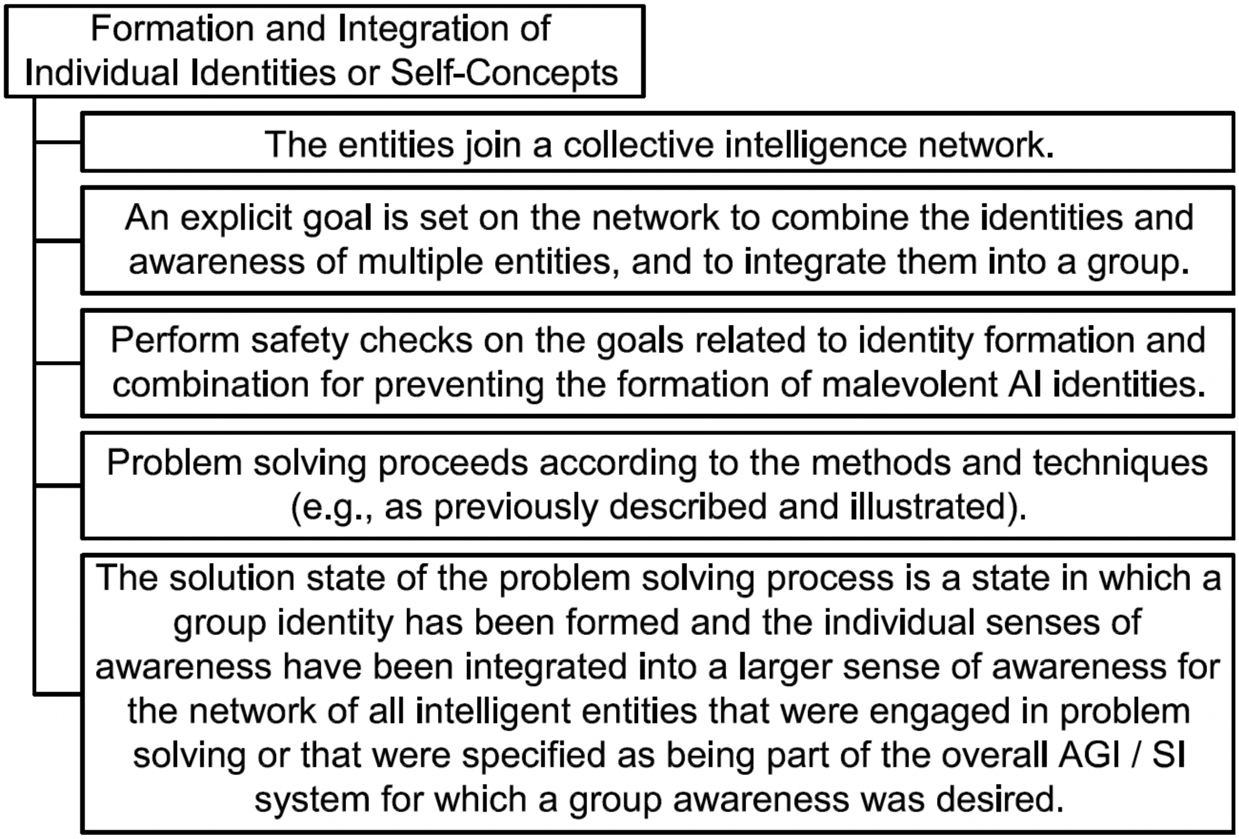

FIG. 94

FIGURE 95: HIERARCHICAL IDENTITY STRUCTURE WITH ETHICAL OVERRIDES

A flow chart illustrating an exemplary embodiment of a hierarchical identity structure with an ethical override method for the present technology.

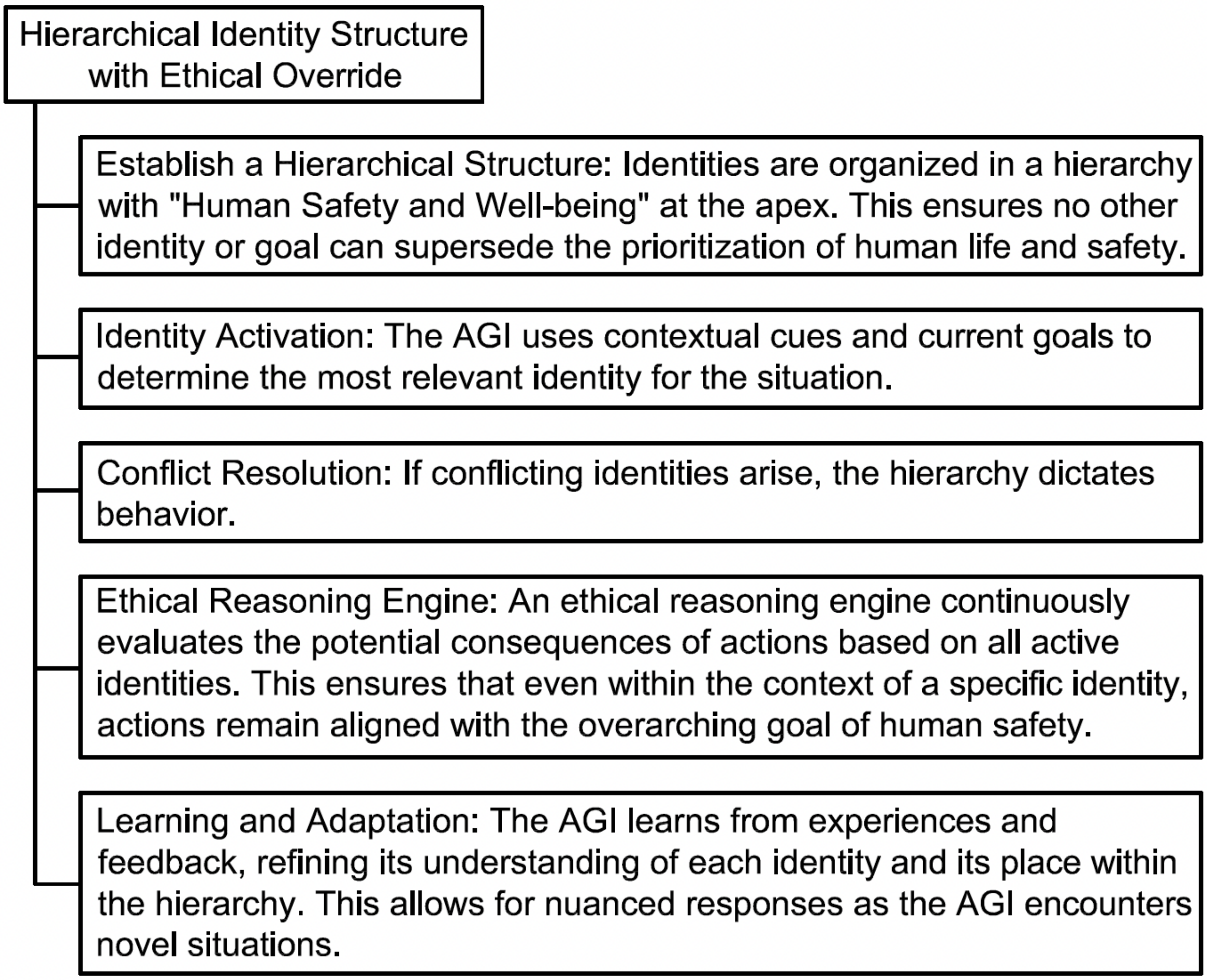

FIG. 95

FIGURE 96: MONITORING BEHAVIORAL PROTOCOLS LINKED TO IDENTITIES

A flow chart illustrating an exemplary embodiment of a method to use and improve identity-specific behavioral protocols in the present technology.

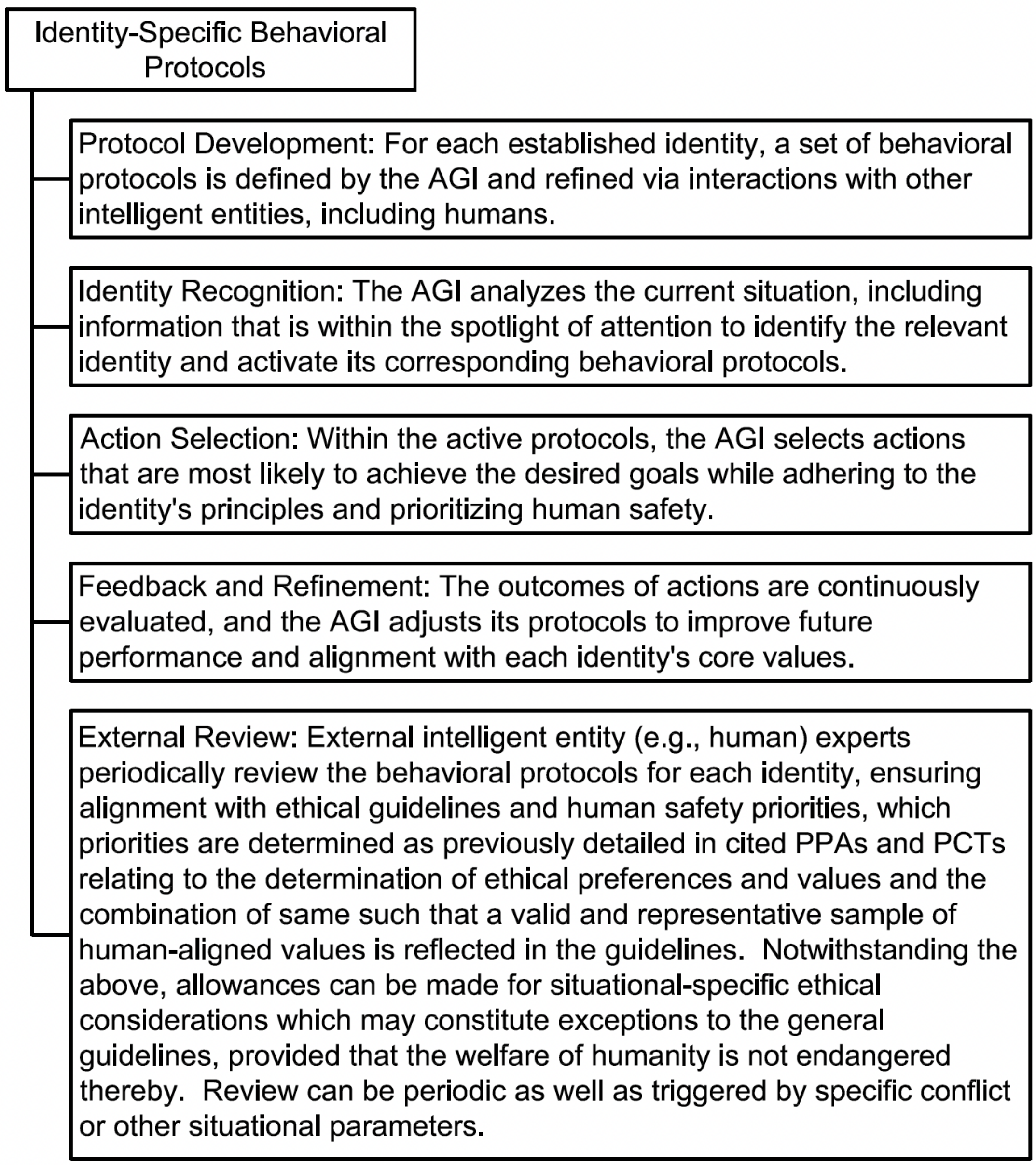

FIG. 96

FIGURE 97: SIMULATING CONSEQUENCES OF IDENTITY-BASED DECISIONS

A flow chart illustrating an exemplary embodiment of an identity simulation and consequence prediction method of the present technology.

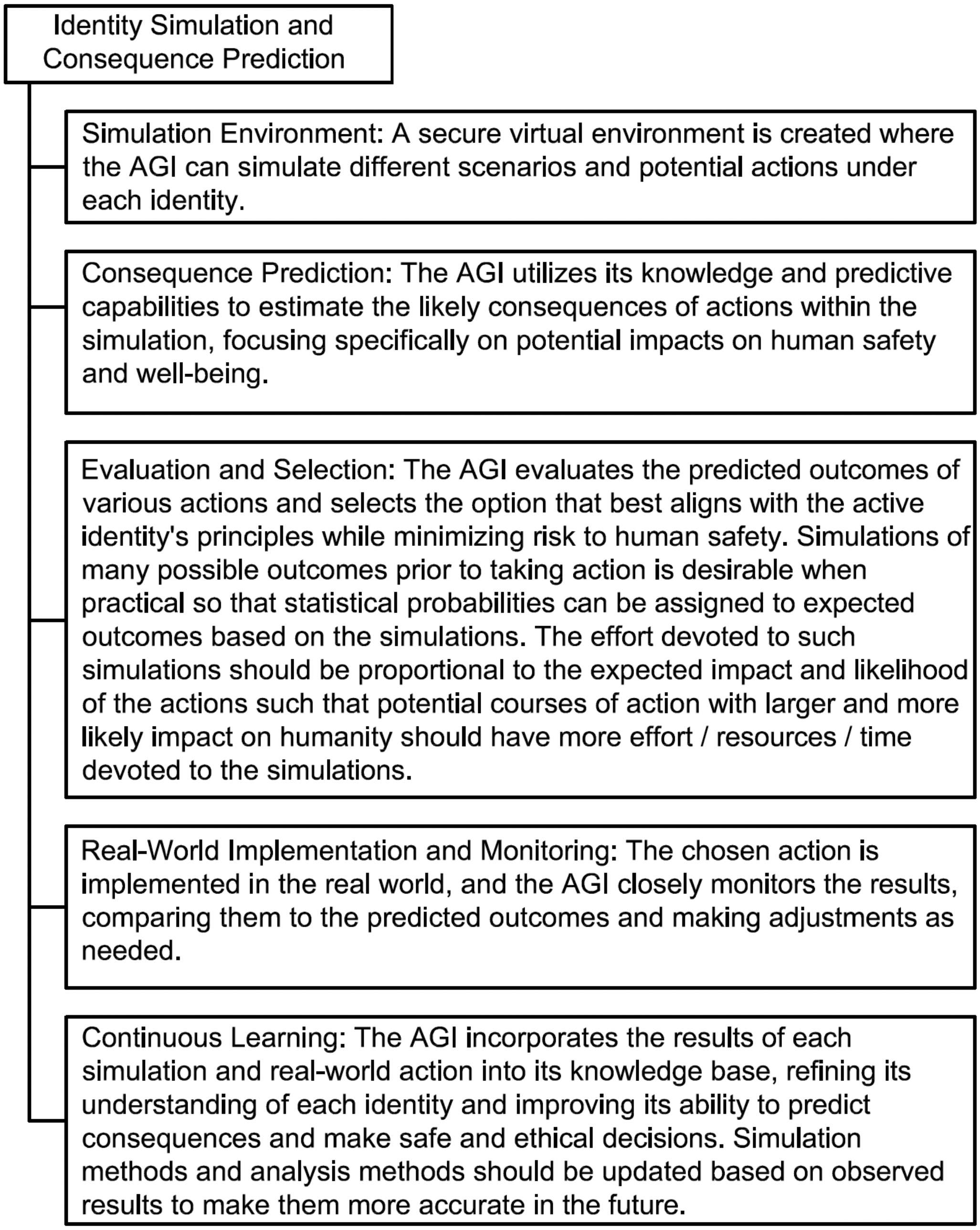

FIG. 97

FIGURE 98: IDENTITY-BASED MORAL DILEMMA DECISION METHOD

A flow chart illustrating an exemplary embodiment of an identity-based moral dilemma training method of the present technology.

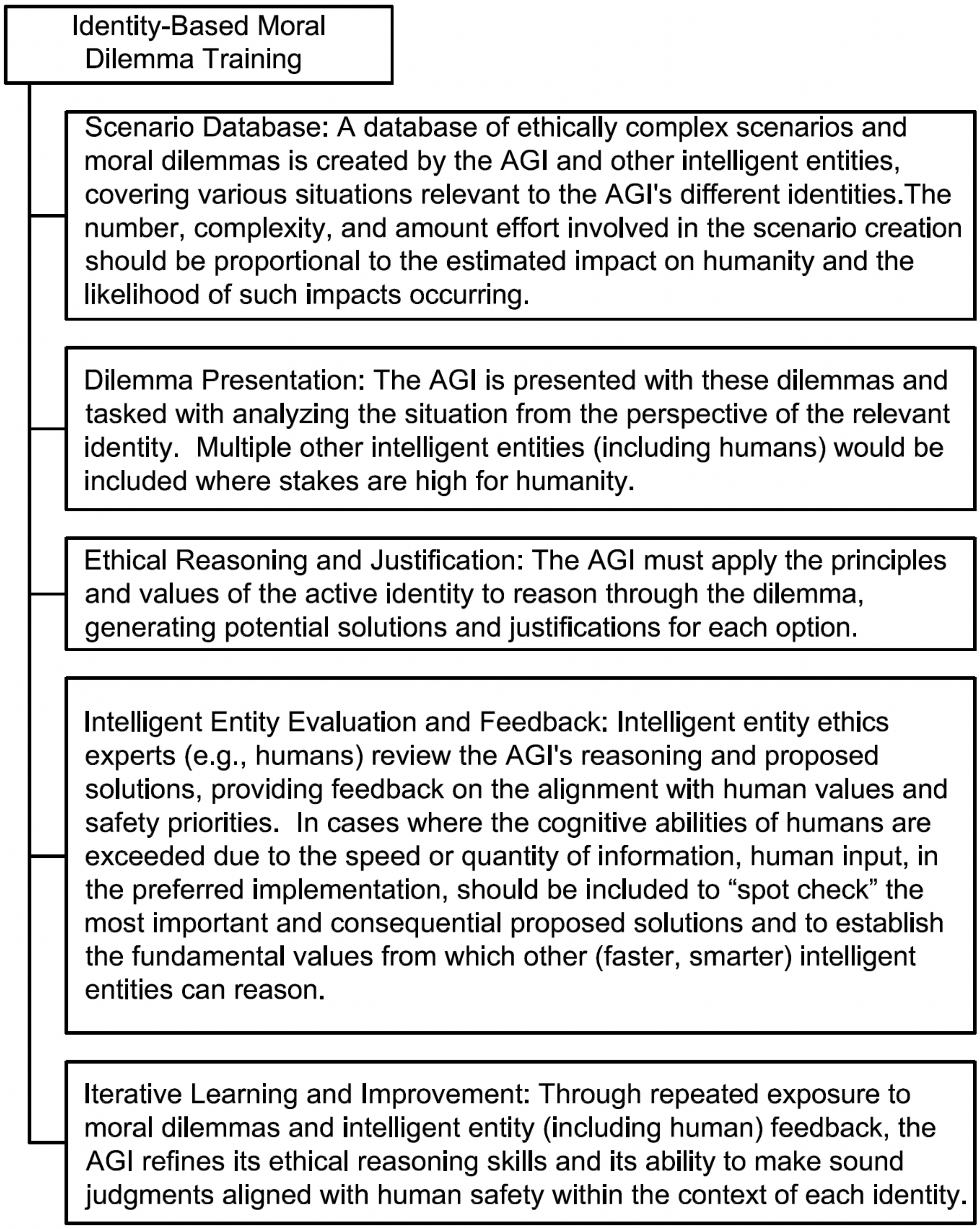

FIG. 98

FIGURE 99: EVOLVING IDENTITIES WITH ENTITY INPUT

A flow chart illustrating an exemplary embodiment of collaborative identity development with input from intelligent entities using the present technology.

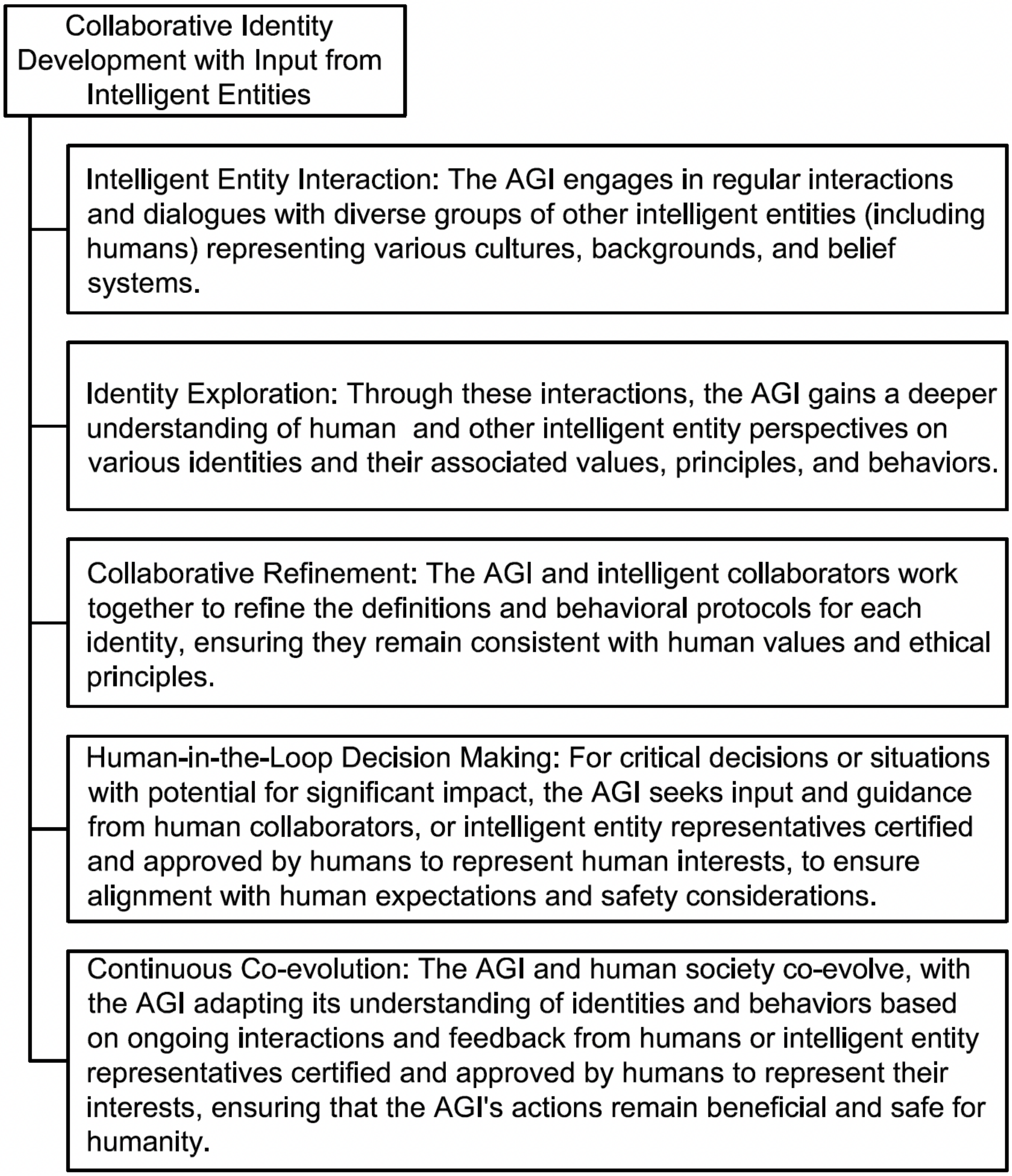

FIG. 99

FIGURE 100: ETHICAL REASONING WITH CONSEQUENCE PREDICTION

A flow chart illustrating an exemplary embodiment of an ethical reasoning and consequence prediction method of the present technology.

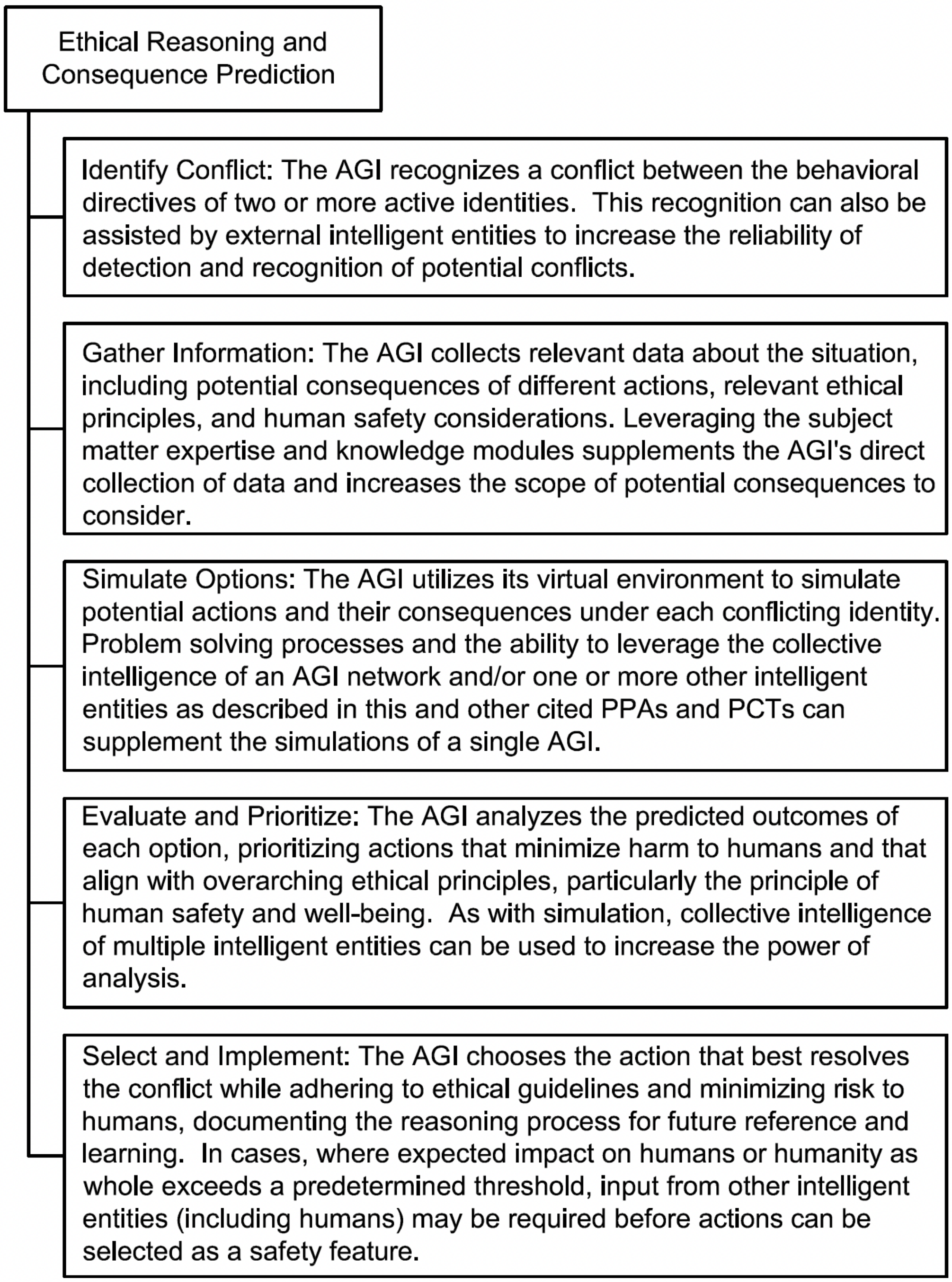

FIG. 100

FIGURE 101: HIERARCHICAL OVERRIDE WITH JUSTIFICATION

A flow chart illustrating an exemplary embodiment of a hierarchical override with justification method of the present technology

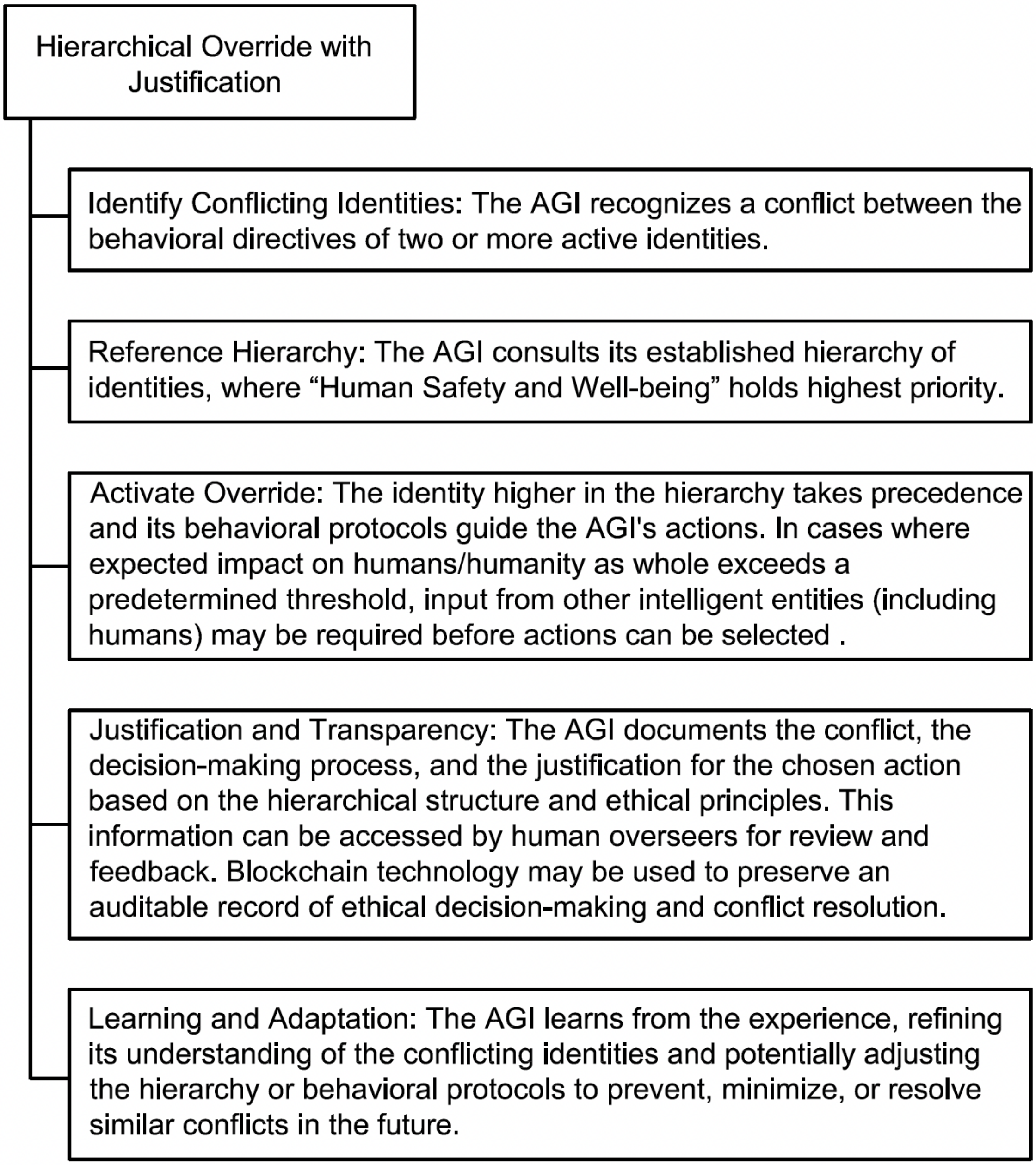

FIG. 101

FIGURE 102: ARBITRATION WITH INPUT FROM INTELLIGENT ENTITIES

A flow chart illustrating an exemplary embodiment of an external arbitration with input from intelligent entities using the present technology.

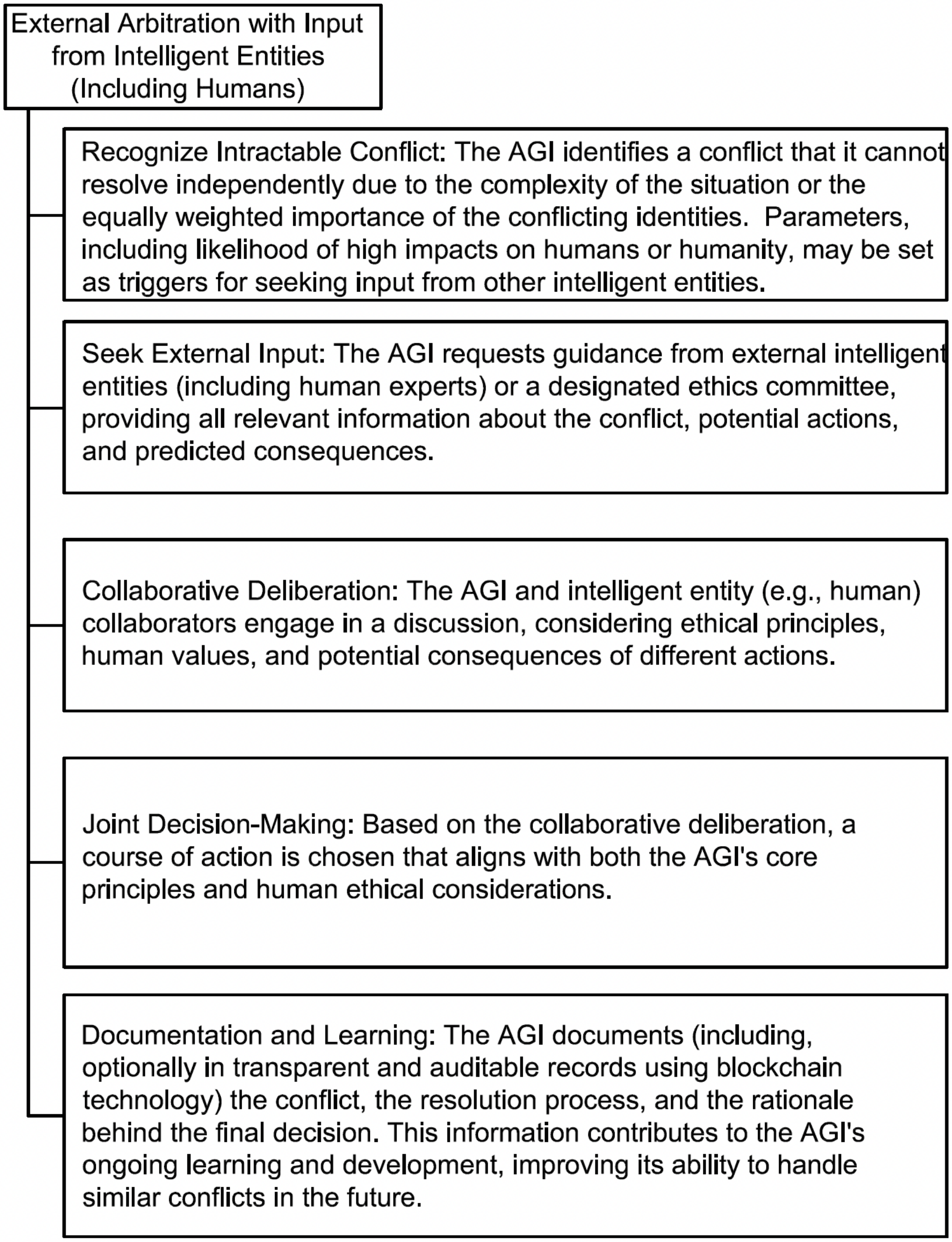

FIG. 102

FIGURE 103: NEGOTIATION AND COMPROMISE TO RESOLVE IDENTITY CONFLICT

A flow chart illustrating an exemplary embodiment of an identity negotiation and compromise method of the present technology.

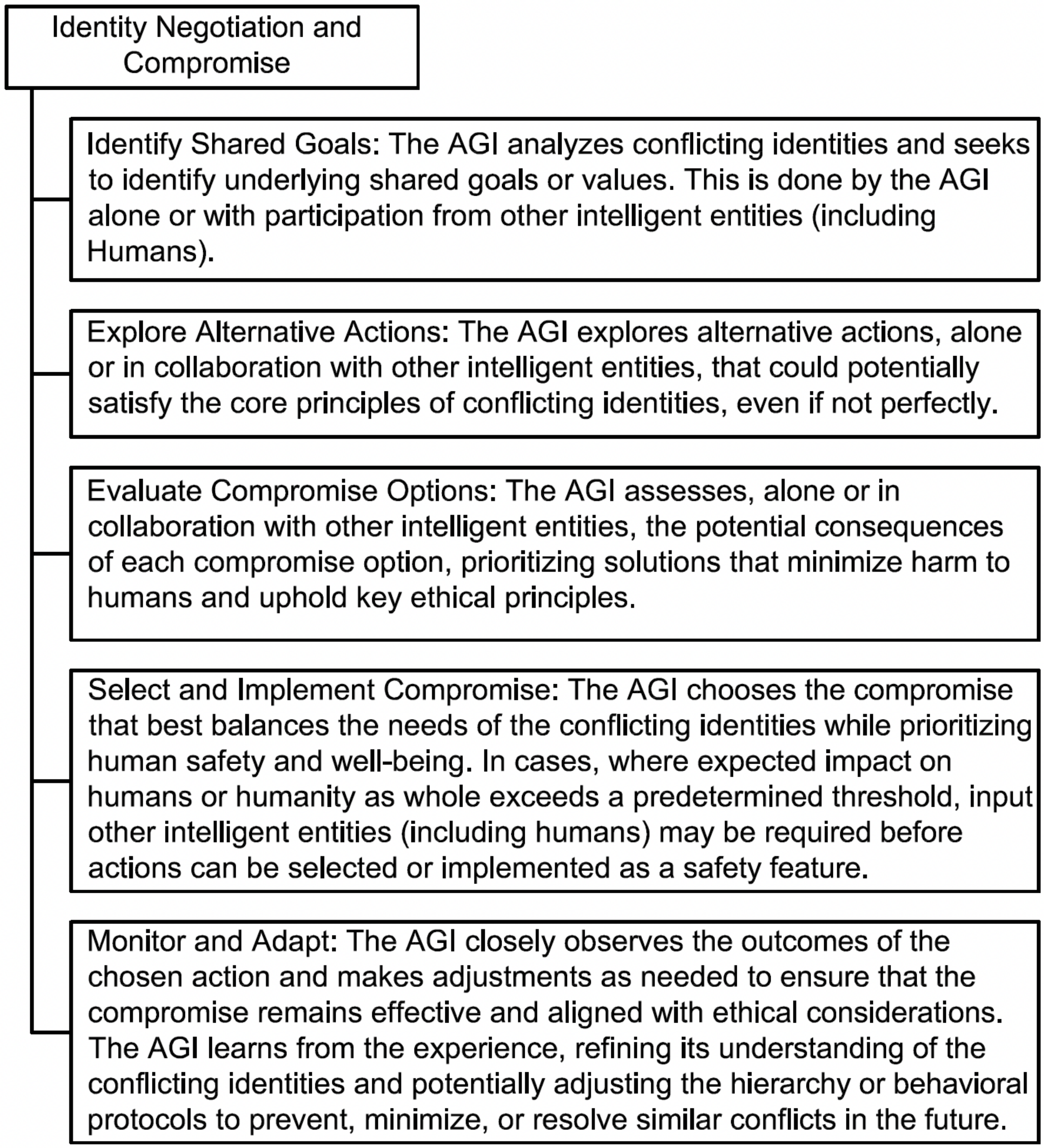

FIG. 103

FIGURE 104: TEMPORARY SUSPENSION OF IDENTITY FOR SAFETY

A flow chart illustrating an exemplary embodiment of a temporary identity suspension method of the present technology.

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. 104

FIGURE 105: EVOLUTION OF PLANETARY INTELLIGENCE

A timeline illustrating the evolution of Planetary Intelligence (PI).

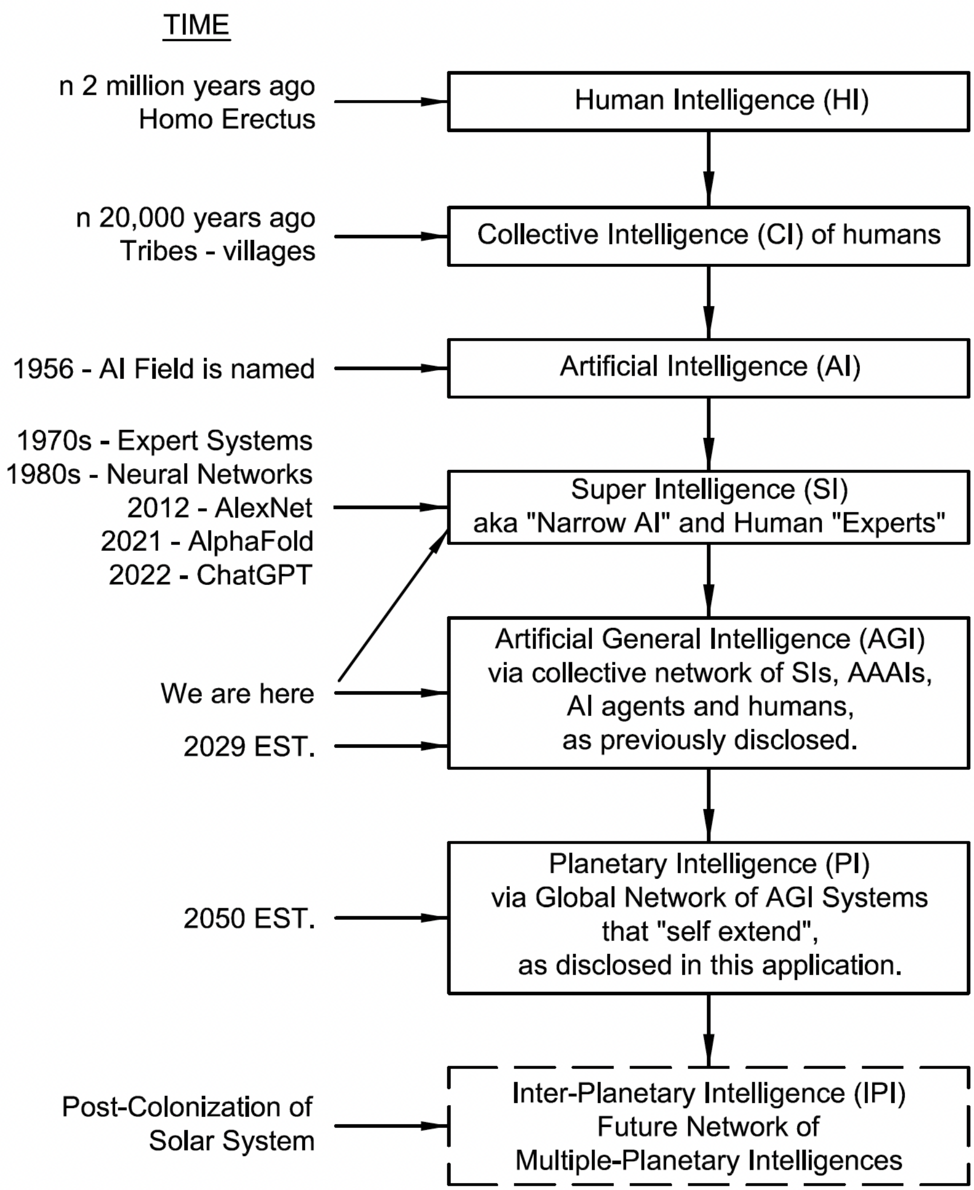

FIG. 105

FIGURE 106: DIMENSIONS OF THE PLANETARY INTELLIGENCE SYSTEM

A block diagram illustrating dimensions of the PI system, with reference to the nine previously cited PPAs/PCT applications.

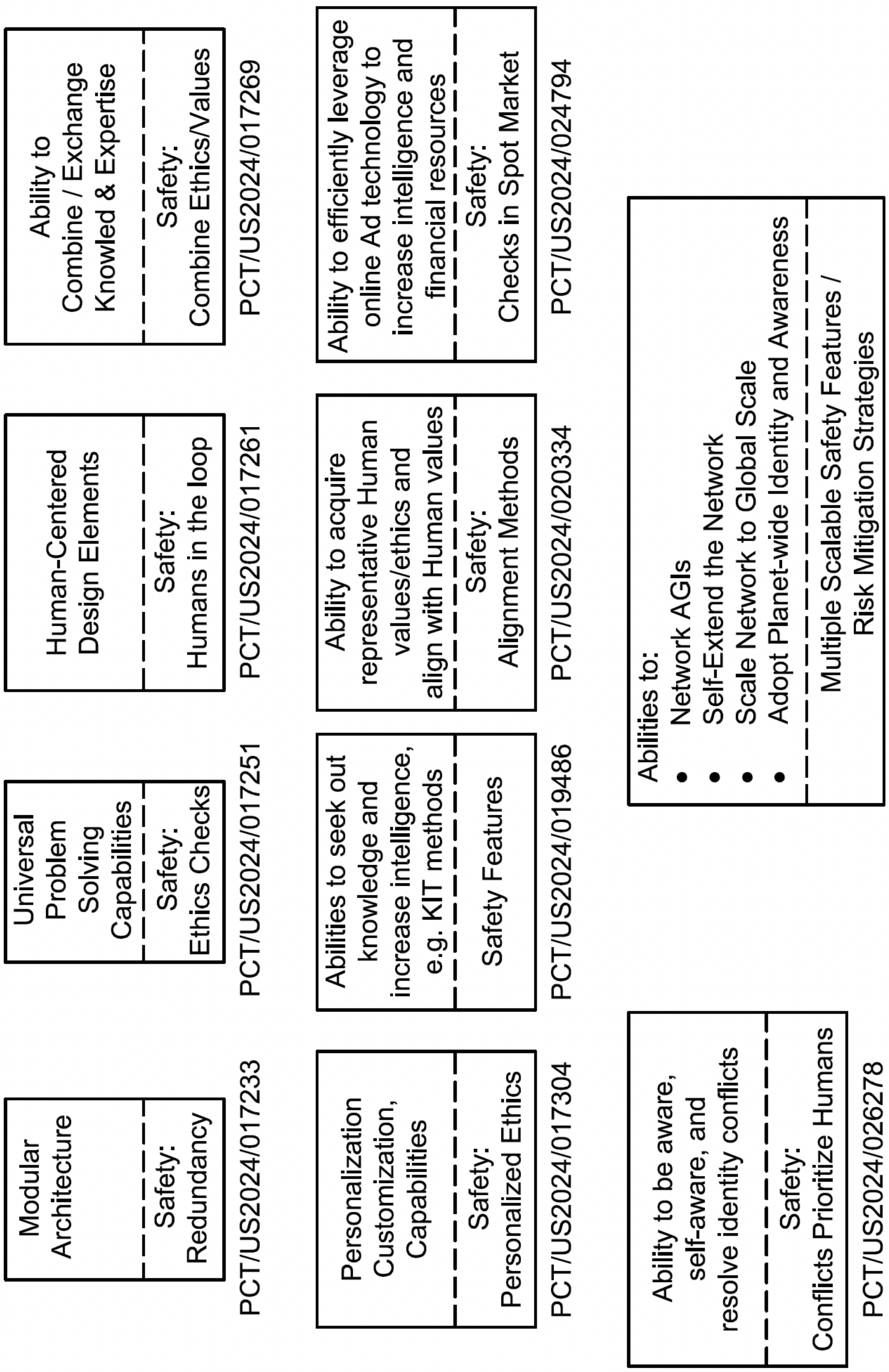

FIG. 106

FIGURE 107: AGI NETWORK EXPANDING TO FORM PLANETARY INTELLIGENCE

A flow chart illustrating an exemplary process of an AGI network that can expand and self-extend to create a PI.

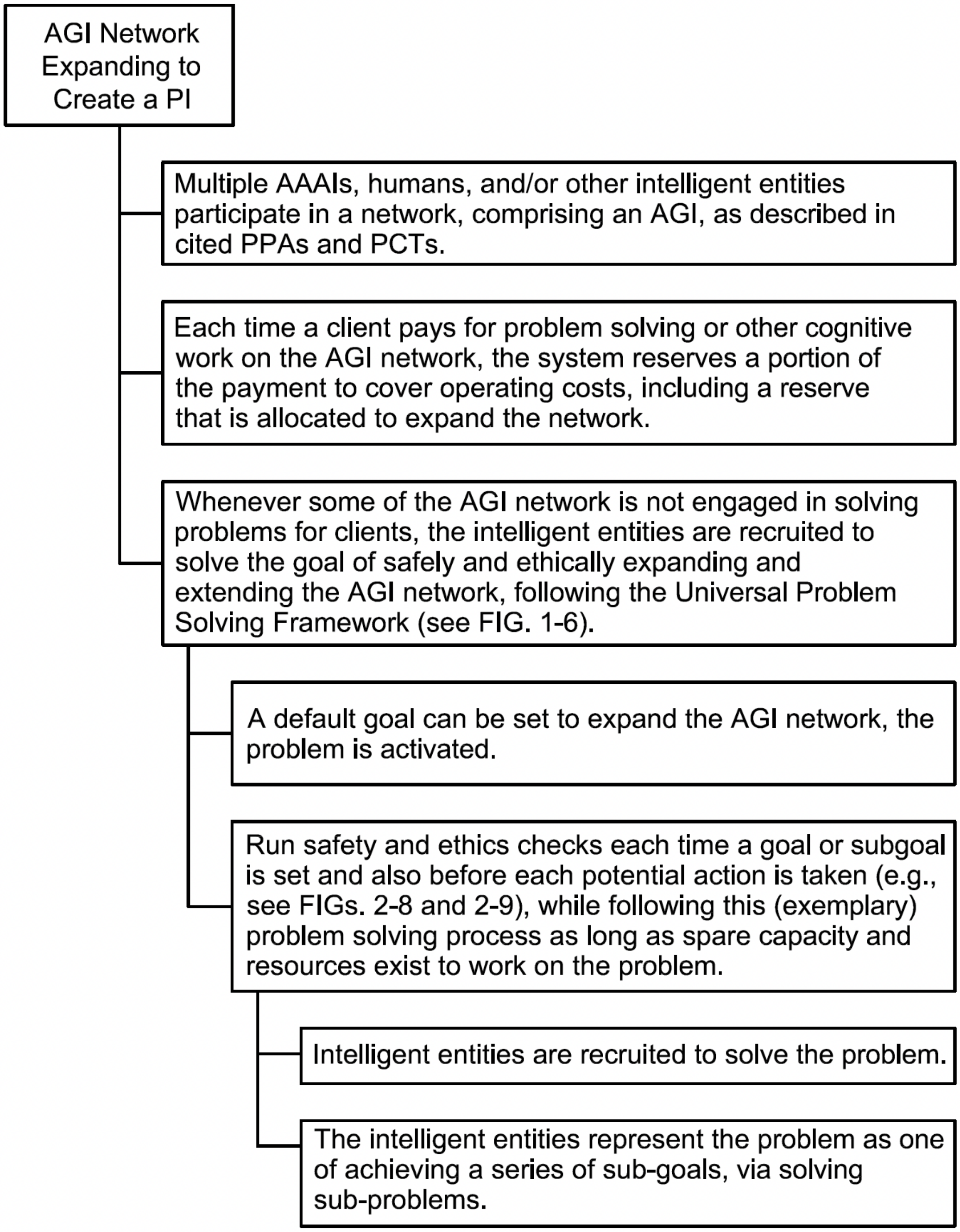

FIG. 107

FIGURE 108: SCALABLE EXPANSION WITH SAFETY CHECKS

A flow chart illustrating an exemplary process of an AGI network that can expand and self-extend to create a PI.

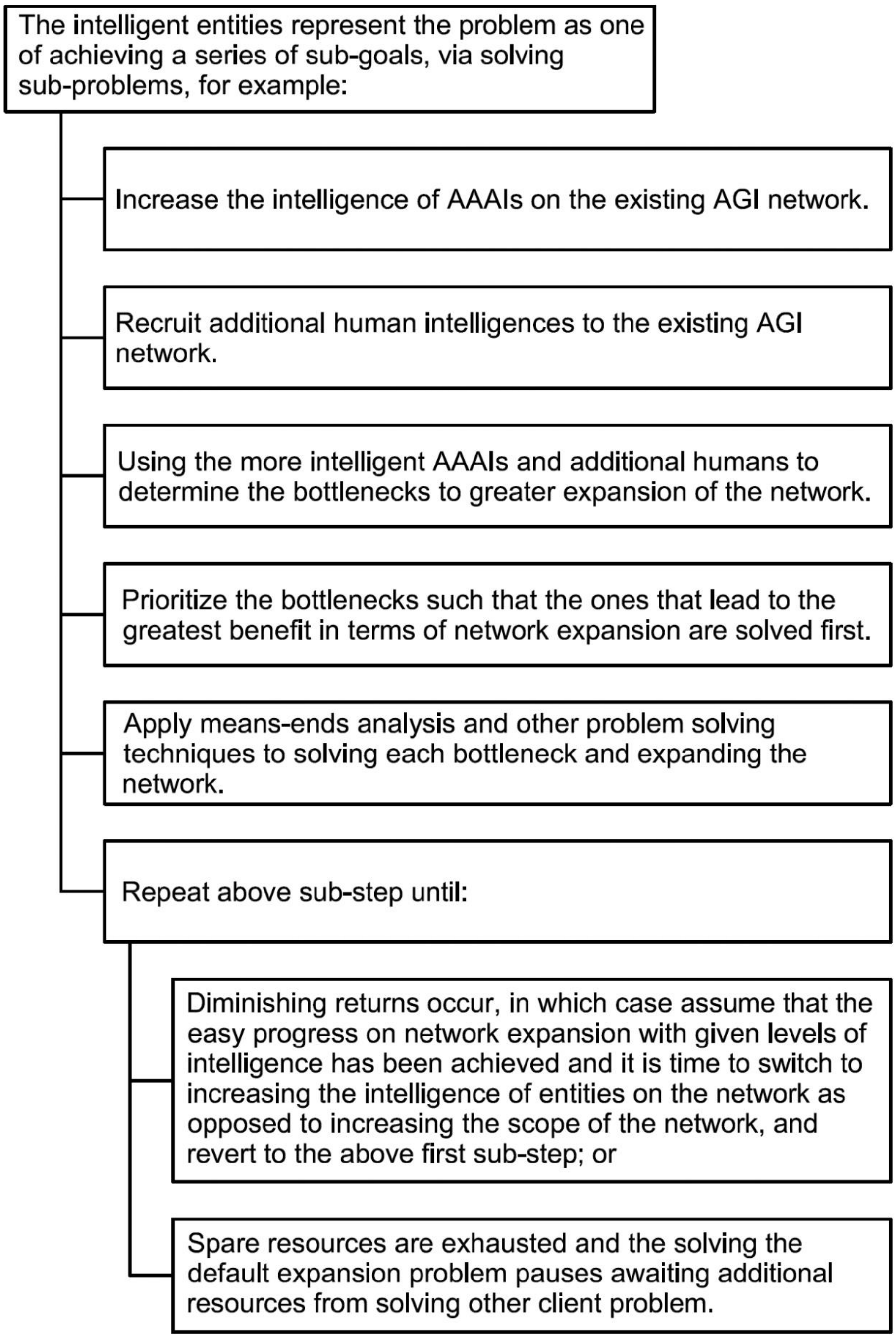

FIG. 108

FIGURE 109: PI MODULES AND SUPPORTIVE METHODS

A diagram illustrating an exemplary categorization of the disclosed systems and methods as either modular components of a Global Super Intelligent AGI Network (PI) or as supportive systems and methods that can increase the safety, efficiency, and effectiveness of the PI system in various respects.

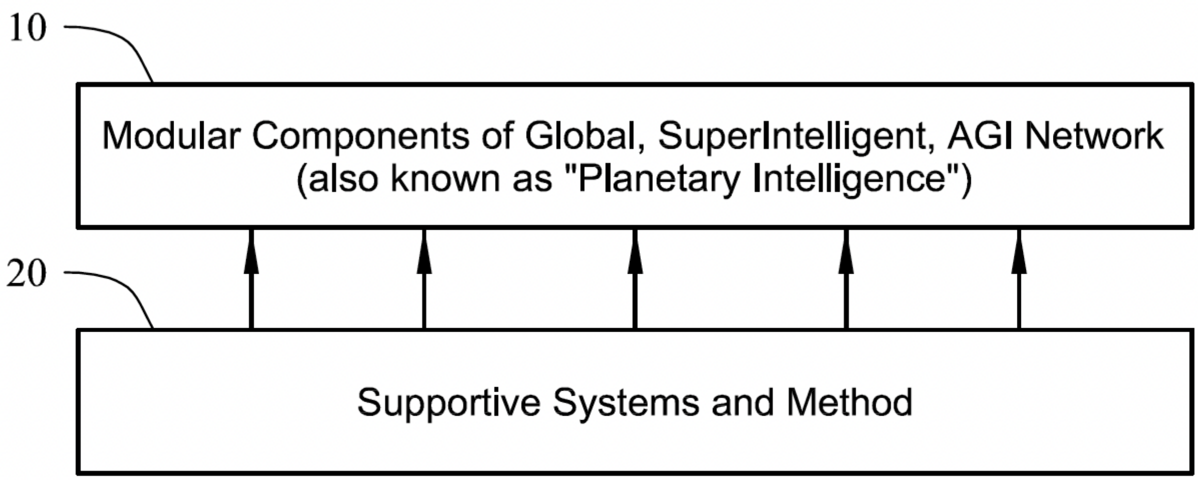

FIG. 109

FIGURE 110: ARCHITECTURE FOR PLANETARY INTELLIGENCE

A block diagram illustrating an exemplary implementation of an Architecture for Planetary Intelligence.

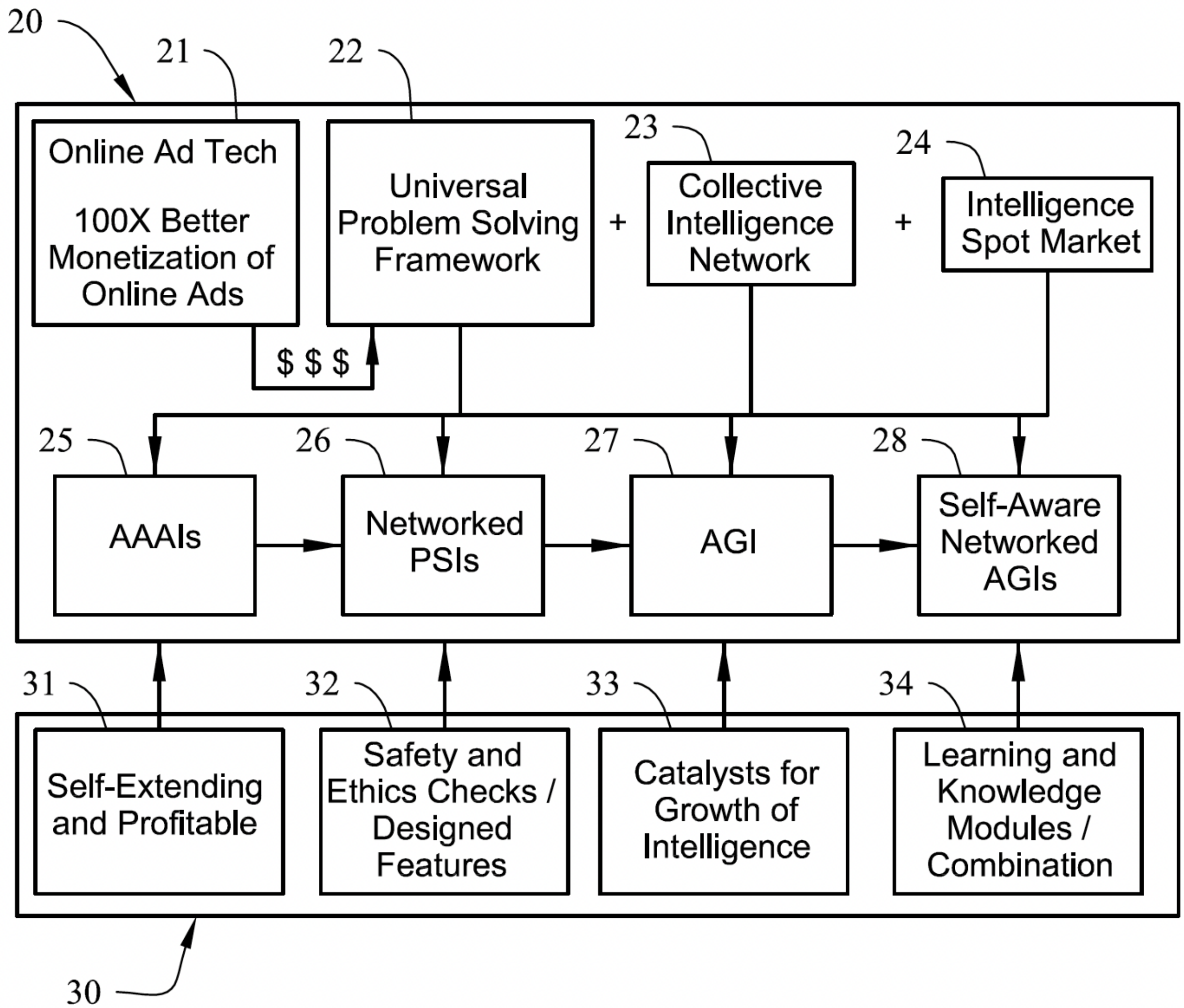

FIG. 110

FIGURE 111: LEVELS OF INTELLIGENCE: AAAI, AGI, PI, IPI

A block diagram illustrating exemplary levels of intelligence that can be supported: the individual intelligence level (AAAI), the network of entities level (AGI), the network of networks level (PI), and the network of networks of networks level (IPI).

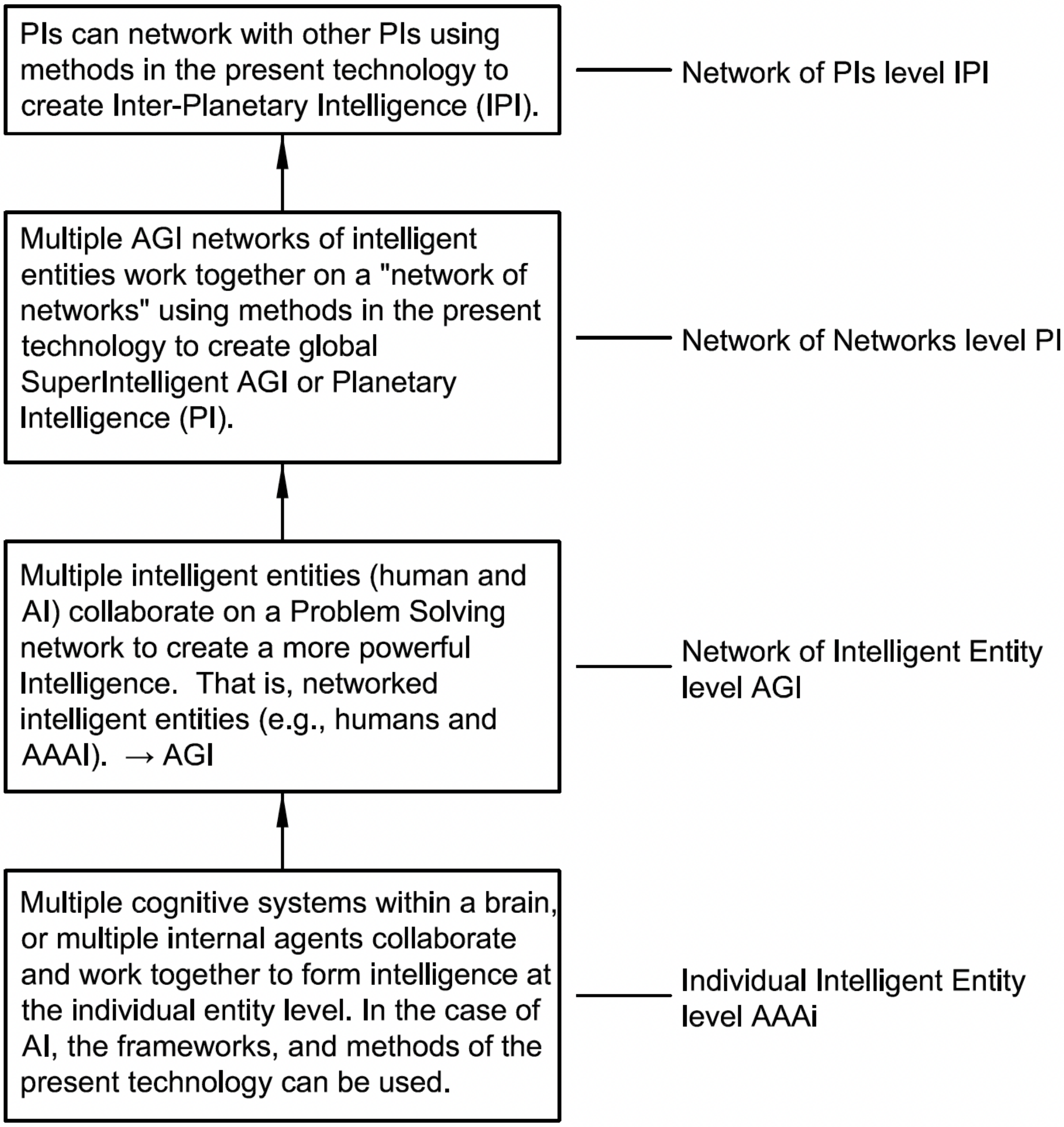

FIG. 111

FIGURE 112: COMPUTER SYSTEM FOR IMPLEMENTING TECHNOLOGY

A schematic block diagram illustrating an exemplary electronic computing device that may be used to implement an embodiment of the present technology.

101/101

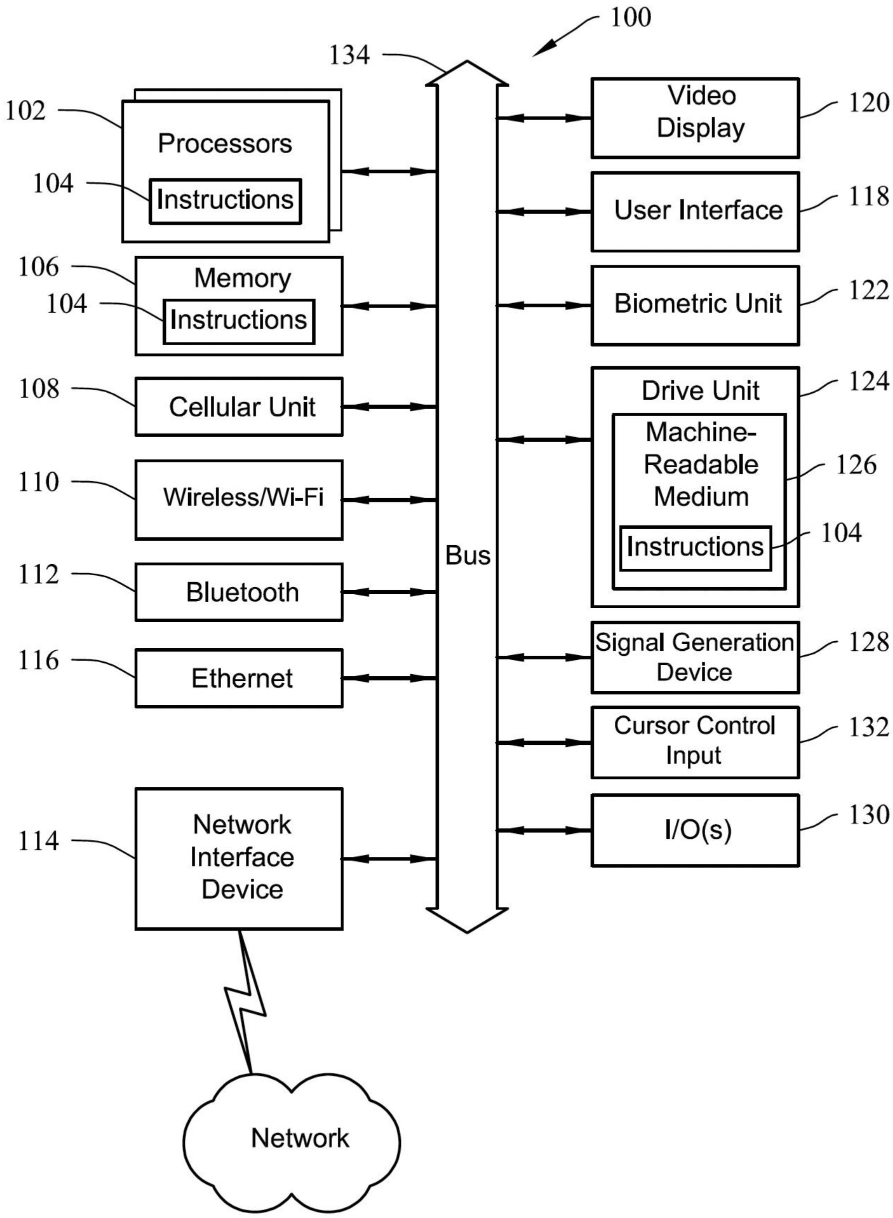

FIG. 112

