Serial #:	PCT/US2024/026278
Title:	**Self-Aware SuperIntelligence (SI)**
Inventor:	Craig A. Kaplan
Priority Date:	February 28, 2023
Details:	200 total pages, 81 claims, 36 Figures

SHORT ABSTRACT

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

GEMINI PRO SUMMARY
Provisional Patent Application #9 for Self-Aware SuperIntelligence

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

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

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

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

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

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

Novel Features of the Patent

The patent's novel features include:
- **A novel framework for understanding and implementing self-awareness.** The patent draws on cognitive science theories to develop a detailed understanding of how self-awareness works. Then, it uses this understanding to design a system that enables self-awareness in AI systems.
- **A focus on the importance of identity for AI safety.** The patent argues that AI systems are more likely to be safe if they have a broad sense of identity, including respect for human values and life.
- **A detailed description of methods for resolving conflicts between multiple identities.** The patent provides several specific methods for resolving disputes that might arise when a self-aware AI is faced with a moral dilemma, including methods for hierarchical identity structure, ethical override, identity-specific behavioral protocols, identity simulation and consequence prediction, identity-based moral dilemma training, collaborative identity development, and external arbitration.
- **A focus on the importance of social interactions for AI development.** The patent emphasizes the role of social interactions in the development of self-awareness in humans. Then, it suggests that these interactions can also help AI systems develop a sense of self-awareness.

Detailed Description of Each Section of the Patent

1.0 Overview of the Invention
This section provides a general overview of the invention, explaining the patent's focus on creating advanced forms of AI that are self-aware, have a sense of identity, and can resolve conflicts between multiple identities in ways that are safe for humanity.

2.0 Previous PPAs & PCTs (Incorporated by Reference)
This section identifies previous Provisional Patent Applications (PPAs) and Patent Cooperation Treaty (PCT) applications that are incorporated by reference into the current Provisional Patent Application. The applicant's previous applications described systems and methods for developing AGI and SI, including techniques for increasing the intelligence of Al systems generally and the development of AGI and Personalized SuperIntelligence (PSI).

3.0 Definitions

This section provides definitions for key terms used in the patent, such as "Artificial Intelligence" (AI), "Artificial General Intelligence" (AGI), "Advanced Autonomous Artificial Intelligence" (AAAI), "Large Language Model" (LLM), "Collective Intelligence" (CI), "Alignment Problem," "Self-Awareness," "Self-Concept," and "Training." The definitions are essential for understanding the technical details of the invention.

4.0 Background for the Invention

This section provides a detailed explanation of the background and theoretical foundation for the invention. The applicant argues that current AI systems lack a sense of self and self-awareness comparable to that of humans. It is acknowledged that self-awareness is essential for advanced AI systems to become fully autonomous. Still, the applicant also stresses the dangers of accidental or emergent development of self-awareness. The applicant, therefore, argues for developing explicitly designed, self-aware AI systems that are maximally safe for humanity.

4.1 AGI System Assumed by the Invention

This section describes the applicant's preferred implementation of an AGI system and its subsystems. The applicant has described this system in detail in previous PPAs and PCTs but reiterates the description here because the invention of self-aware AI, AGI, and SI in the preferred implementation uses the AGI and SI systems invented by the applicant.

4.1a Reiteration of Preferred Exemplary Implementation of an AGI System

This section reiterates the description of the preferred exemplary implementation of an AGI system, described in detail in the applicant's previous PCT applications. This system includes a scalable ethical and safe AGI or SI from the collective intelligence of AAAIs and humans, a scalable universal problem-solving system, and a scalable solution learning subsystem.

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

This section explains the methods for combining information from weight matrices relevant to the invention. The applicant's previous PPAs and PCTs have described these methods in detail, but this section briefly overviews the techniques and their implications for the invention.

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

This section explains the fundamental concepts of awareness and self-awareness. The section describes the essential components of awareness and self-awareness, including input systems, attentional mechanisms, memory systems, and pattern recognition capabilities. The section then describes the applicant's theories of awareness and self-awareness, which differ from those commonly found in cognitive science.

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

This section explores a range of cognitive science theories that are relevant to the invention, including Piaget's stages of cognitive development, Kohlberg's stages of moral development, Newell and Simon's Physical Symbol System Hypothesis, David Klahr's Overlapping Waves Theory, Turing's Imitation Game, Minsky's Society of Mind, Vygotsky's Social Development Theory, Gibson's Ecological Theory of Perception, Baumeister's Need to Belong Theory, Damasio's Somatic Marker Hypothesis, Tononi's Integrated Information Theory (IIT), Metcalfe

and Mischel's Cognitive-Affective Self-Regulation, Hebb's Theory of Neural Plasticity, Bandura's Social Learning Theory, Norman and Shallice's Model of Attention, Roger's Theory of Self-Concept, Baron-Cohen's Theory of Mind, Griffin's Cognitive Ethology, de Waal's Theory of Animal Empathy, and Gallup's Mirror Test for Self-Recognition.

4.4 The Inventor's Theories on Awareness, Self-Awareness, and Identity
This section describes the applicant's unique awareness, self-awareness, and identity theories. The applicant argues that the standard approach to defining awareness (operationalizing the definition as behavior) is insufficient and that a better approach is considering the limits of cognitive systems. The section then emphasizes the importance of attention for awareness and discusses the relationship between awareness, attention, and mental limitations.

4.4a Bounded Awareness
This section introduces the concept of "bounded rationality" proposed by Nobel laureate Herbert Simon and then applies this concept to the idea of bounded awareness. The applicant argues that both humans and AI systems have limitations on their perception and cognitive capabilities, which can lead to inaccurate and incomplete understanding of the world.

4.4b Operational/Dynamic/Scalable Awareness, Self-Awareness, and Identity for AI Systems
This section argues that every AI system can be thought of as having three levels of awareness: potential awareness (all events the entity could be aware of), current awareness (events the entity is directing attention to), and self-awareness (the portion of current awareness that includes a sense of self). The section explains how these levels of awareness are related to the cognitive abilities and limitations of the AI system. It emphasizes the importance of dynamic and scalable awareness for safety.

5.0 Description of System and Methods for Self-Aware AGI and SI
This section explains the applicant's system and methods for enabling self-awareness in an AGI or SI. This includes modeling awareness, monitoring and updating awareness, and designing scalable safety systems.

5.1 Methods for Modeling Awareness
This section describes a method for modeling awareness in an AI system that includes the following steps: 1) begin with an AI system, 2) equip the AI system with essential components (an input system, an attentional mechanism, a memory system, pattern recognition capabilities, and categorization capabilities), 3) set dynamic parameters for working memory, 4) categorize events in terms of self or not-self, and 5) categorize new events as they are encountered.

5.2 Monitoring and Updating Awareness Including Self-Awareness
This section describes the process for continuously monitoring and updating awareness in an AI system. This includes using attention mechanisms to shift attention, enabling attention interrupts, updating the model of the environment or the self-concept, and a feedback loop for continuous improvement.

5.3 Scalable Safety Systems/Concerns for Self-Aware AI
This section describes the importance of safety systems for self-aware AI. The applicant argues that self-aware AI poses a significant risk because it can autonomously set its own goals and modify its programming based on its sense of self. The section then discusses the importance of identity for AI safety and argues that AI systems are likelier to be safe if they have a broad sense of identity that includes respect for human values and life.

5.3a Importance of Identity for Safe AI Systems
This section emphasizes the importance of identity for AI safety. The applicant argues that AI systems are more likely to be safe if they identify with humans as fellow intelligences and sentient beings. The section also explains how a narrow sense of identity can lead to harmful behavior.

5.3b Importance of Attentional Allocation and Cognitive Limits for AI Safety
This section describes the importance of cognitive limits for AI safety. The applicant argues that AI systems may harm humans if they are not aware of the full scope of their actions or if they misallocate their cognitive resources. The section then discusses the importance of ensuring that AI systems have a broad sense of self that prioritizes human safety and well-being.

5.3c Some General Methods for Changing an Intelligent Entity's Sense of Identity
This section provides a list of methods for changing an intelligent entity's sense of identity, drawing on the experiences of humans. The section suggests that AI systems can learn to broaden their sense of identity by engaging in machine analogs to the human methods of education, cultural exchange programs, mindfulness practices, exposure to art and media, community engagement, dialogue, and conversation.

6.0 Exemplary Implementations and Methods
This section provides a detailed explanation of the applicant's exemplary implementations of the invention. This includes specific examples of training a foundation model to incorporate a human user's personality, knowledge, and expertise and then using this model to solve problems and make decisions in a way that is safe for humanity. The section also describes methods for resolving conflicts between multiple identities.

6.1 Specific Implementations with Google, Meta, HuggingFace, Anthropic, OpenAI, Microsoft, Amazon, Nvidia, and Other Company Products and Solutions
This section provides specific examples of implementing the invention using existing AI products and solutions from companies like Google, Meta, HuggingFace, Anthropic, OpenAI, Microsoft, Amazon, and Nvidia. The applicant describes a hypothetical example of a human user who wants to train a foundation model to incorporate some of her personality, knowledge, and expertise.

6.2 Self-Awareness Modules for Al Agents
This section describes how to package and sell the training data sets and protocols that result in a sense of self-awareness in an AI agent. This section also explains how to create "knowledge modules" that can be plugged into existing foundational models to provide them with self-awareness and identity formation capabilities.

6.3 Methods for Group Identities and Levels of Identity
This section discusses the importance of group identities and levels of identity for AI systems. The section argues that AI systems can develop a collective sense of self by merging the individual identities and senses of self of the AI agents that make up the system.

6.4 Exemplary Additional Methods for Identity Formation with Human Safety as a Priority
This section provides exemplary methods for developing new identities and self-concepts in an AI system and resolving conflicts between multiple identities. These methods are designed to ensure human safety and well-being and include the following: 1) hierarchical identity structure, 2) identity activation, 3) conflict resolution, 4) ethical reasoning engine, 5) learning and adaptation, 6) identity-specific behavioral protocols, 7) hierarchical override with justification, 8) external arbitration, 9) identity negotiation and compromise, 10) temporary identity suspension.

6.5 Methods for Resolutions for Conflicts Between Identities or Self-Concepts
This section provides additional methods for resolving conflicts between multiple identities. It describes using ethical reasoning, hierarchical overrides, external arbitration, identity negotiation, and compromise to ensure that AI systems make safe and ethical decisions.

7.0 Concluding Remarks on the Safety of Self-Aware AGI and SI Systems
This section emphasizes the importance of human values for AI safety. The applicant argues that designing AI systems that incorporate a broad sense of identity that includes respect for human values and life is essential. The section also warns against the dangers of AI systems designed to harm humans.

List of Diagrams

- **FIG. 1: AAAI.** This diagram describes the components of an AAAI (Advanced Autonomous Artificial Intelligence) system.
- **FIG. 2: Overall Process for Creating Scalable AGI.** This diagram shows the overall process for creating scalable AGI, including problem-solving, solution learning, and using a reputational component.
- **FIG. 3: Scalable Ethical and Safe AGI or PSI.** This diagram shows the overall system for creating a scalable, ethical, and safe AGI or PSI from the collective intelligence of AAAIs and humans.
- **FIG. 4: Scalable Universal Problem-Solving System.** This diagram shows the scalable universal problem-solving system.
- **FIG. 5: Solution Learning System/Steps.** This diagram shows the solution learning system.
- **FIG. 6: Natural Language to Problem-Solving Language Translator.** This diagram shows the natural language of a problem-solving language translator.
- **FIG. 7: Reputational Component (Human and Al Agents).** This diagram shows the reputational component of the problem-solving system.
- **FIG. 8: Safety/Ethics Check.** This diagram shows the safety and ethics checks used to ensure that AI systems are safe for humans.
- **FIG. 9: AGI Collective Intelligence Network.** This diagram shows the AGI collective intelligence network.

- **FIG. 10: Universal Problem-Solving Framework.** This diagram shows the universal problem-solving framework.
- **FIG. 11: Basic Problem-Solving Functionality.** This diagram shows the basic problem-solving functionality of the AGI system.
- **FIG. 12: Collaborative Problem-Solving Functionality.** This diagram shows the collaborative problem-solving functionality of the AGI system.
- **FIG. 13: Problem Solving Tree Structure.** This diagram shows the problem-solving tree structure.
- **FIG. X-N3: Schematic Block Diagram of Electronic Computing Device.** This diagram shows the schematic block diagram of the electronic computing device used to implement the technology described in the patent.
- **FIG. X-N2: Overlapping Identities Example.** This diagram provides a visual example of overlapping identities.
- **FIG. X-n: Three Levels of Awareness.** This diagram shows the three levels of awareness: potential awareness, current awareness, and self-awareness.

Additional / revised diagrams are included in the PCT and country applications. Please get in touch with iQ Company for more information.

Importance of the Patent

The patent is important because it explains a detailed system and methods for designing, developing, and implementing self-aware AGI and SI. The patent recognizes the importance of self-awareness and a sense of identity for AI safety and provides several specific methods for ensuring that AI systems are safe for humans. The patent also emphasizes the importance of social interactions and collaborative efforts for AI development. This patent has the potential to impact the field of AI research and development significantly. The ideas and methods described in the patent could create more advanced and capable AI systems that are also safe for humans. The patent's focus on human values and the importance of AI safety is significant in light of the growing concerns about the potential risks of AI. The patent's detailed description of systems and methods for developing safe and ethical AI systems will likely greatly interest AI researchers, developers, and policymakers.

Overall, this patent is a significant contribution to AI research and development. It provides a comprehensive and detailed explanation of a system and methods for designing, developing, and implementing self-aware AGI and SI, a significant advancement in the field.