SUPERINTELLIGENCE
AN iQ COMPANY

ABSTRACT & SUMMARY SUPERINTELLIGENCE DESIGN WHITE PAPER #9: SELF-AWARE SUPERINTELLIGENCE

by Dr. Craig A. Kaplan
May 2025

ABSTRACT

This white paper describes how to add the dimension of self-awareness and increased autonomy to the AI, AGI, and SuperIntelligent systems described in previous white papers. We present inventions related to: attention, attentional interrupts, modeling and maintaining awareness and self-awareness, training and tuning of models, novel versions of the Turing Test, forming individual and group identities, combining identities, multiple ways (including hierarchical methods) for resolving conflicts between identities, temporary suspension of identities in unsafe conditions, continuous improvement and learning, and other methods that enable AI, AGI, and SI systems to become self-aware and to function with a sense of identity.

Properly implemented, self-aware SuperIntelligence could be the most positive invention in human history. Poorly implemented, it could become the most dangerous.

Therefore, we explain in depth how to design safety in the systems, prevent bad outcomes, and maximize alignment with human values.

SUMMARY

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

The white paper acknowledges that current AI systems lack self-awareness but argues that it is inevitable that advanced AI systems will develop such capabilities. The design addresses this

challenge by detailing a system enabling self-awareness and a sense of identity in an AI/AGI/SI. The system design is based on carefully studying human cognitive systems, including the relationship between awareness, attention, and memory. The author argues that since self-awareness is a special case of general awareness (where the objects of awareness are self and not-self), a system capable of general awareness can be extended to become self-aware.

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

- The importance of a hierarchical identity structure in which human safety is prioritized.
- Using ethical reasoning engines ensures that AI systems' actions align with human values.
- Developing robust feedback mechanisms allows AI systems to learn from their interactions with humans and other intelligent entities.
- There is a need for ongoing training and education in ethics and social norms for AI systems.

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

Novel Features of the White Paper

- **A novel framework for understanding and implementing self-awareness.** The white paper 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 white paper argues that AI systems are more likely to be safe if they have a broad sense of identity that includes a respect for human values and life.
- **A detailed description of methods for resolving conflicts between multiple identities.** The white paper provides several specific methods for resolving conflicts 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, 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 white paper emphasizes the role of social interactions in developing 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 White Paper

Overview of the Design: This section provides a general overview of the design, explaining the white paper'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.

Previous White Papers: This section identifies previous white papers upon which this builds. These white papers describe systems and methods for developing AGI and SI, including techniques for increasing the intelligence of AI systems generally, and the development of AGI and Personalized SuperIntelligence (PSI).

Definitions: This section provides definitions for key terms used in the white paper, 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 design.

Background for the Design: This section provides a detailed explanation of the background and theoretical foundation for the design. It 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 author also stresses the dangers of accidental or emergent development of self-awareness. The author, therefore, argues for developing explicitly designed, self-aware AI systems that are maximally safe for humanity.

AGI System Assumed by the Design: This section describes a preferred implementation of an AGI system and its subsystems. The author has described this system in detail in previous PPAs and PCTs, but reiterates the description because the design of self-aware AI, AGI, and SI in the preferred implementation uses these AGI and SI systems.

Reiteration of Preferred Exemplary Implementation of an AGI System: This section reiterates the preferred exemplary implementation of an AGI system, previously described in

other white papers. 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.

Reiteration of Some Methods for Combining Information from Weight Matrices: This section explains the methods for combining information from weight matrices relevant to the design. Previous white papers have described these methods in detail, but this section briefly overviews them and their implications for the design.

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 author's theories of awareness and self-awareness, which differ from those commonly found in cognitive science.

Cognitive Theories, Related to Developing Self-Awareness in AI Systems: This section explores a range of cognitive science theories that are relevant to the design, 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.

New Theories on Awareness, Self-Awareness, and Identity: This section describes the author's unique awareness, self-awareness, and identity theories. The author argues that the

standard approach to defining awareness (operationalizing the definition as behavior) is insufficient, and that a better approach is to consider the limits of cognitive systems. The section then emphasizes the importance of attention for awareness and discusses the relationship between awareness, attention, and cognitive limitations.

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 white paper 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.

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). This 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.

Description of System and Methods for Self-Aware AGI and SI: This section explains the 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.

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.

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.

Scalable Safety Systems/Concerns for Self-Aware AI: This section describes the importance of safety systems for self-aware AI. The white paper 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. This section emphasizes the importance of identity for AI safety and argues that AI systems are more likely to be safe if they have a broad sense of identity that includes a respect for human values and life.

Importance of Identity for Safe AI Systems: This section emphasizes the importance of identity for AI safety. The author argues that AI systems are likelier 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.

Importance of Attentional Allocation and Cognitive Limits for AI Safety: This section describes the importance of cognitive limits for AI safety. It argues that AI systems may harm humans if they are unaware of the full scope of their actions or 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.

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.

Exemplary Implementations and Methods: This section provides exemplary implementations of the design. 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 safely make decisions for humanity. The section also describes methods for resolving conflicts between multiple identities.

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

Self-Awareness Modules for AI 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.

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.

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, and
10. temporary identity suspension.

Methods for Resolutions of 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.

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

Figures: White Paper #9 contains 34 figures that are described in detail in White Paper #10, Planetary Intelligence.

Importance of the White Paper

- It explains a system and methods for designing, developing, and implementing self-aware AGI and SI.
- It recognizes the importance of self-awareness and a sense of identity for AI safety.

- It provides several specific methods for ensuring that AI systems are safe for humans.
- It emphasizes the importance of social interactions and collaborative efforts for AI development.
- It can impact AI research and development significantly.
- The ideas and methods described could create more advanced and capable AI systems that are also safe for humans.
- The 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 detailed description of systems and methods for developing safe and ethical AI systems will greatly interest AI researchers, developers, and policymakers.

Overall, White Paper #9 significantly contributes 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.

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.

