SUPERINTELLIGENCE
AN iQ COMPANY

ABSTRACT & SUMMARY SUPERINTELLIGENCE DESIGN WHITE PAPER #1: ADVANCED AUTONOMOUS ARTIFICIAL INTELLIGENCE

by Dr. Craig A. Kaplan

May 2025

ABSTRACT

Artificial General Intelligence (AGI) will be the world's most powerful invention. AGI will be able to solve any problem better and faster than humans, or it can make humanity extinct. Whether AGI represents a multi-trillion-dollar opportunity or an existential threat depends critically on AGI's design.

This white paper describes achieving safe AGI by relying on the combined knowledge and ethics of many humans, each of whom customizes their own cloneable Advanced Autonomous Artificial Intelligence, or AAAI. The AAAIs can be customized via a single click, leveraging existing data from social media and other sources of information. These AAAIs collaborate with other humans on a network, using a novel universal problem-solving architecture to comprise an AGI system.

The AGI system is also designed with scalable safety features integral to the system architecture. The white paper describes not only the fastest but also the safest path to AGI.

SUMMARY

Design White Paper #1 describes a design for developing Artificial General Intelligence (AGI) and SuperIntelligent AGI that leverages the collective intelligence of millions of humans and AI agents. The design achieves a faster and safer path to AGI by relying, at least initially, on the involvement of humans in the AGI training, operation, and safety/supervisory functions. The design enables users to customize their AI agents (AAAIs) and then have those AI agents participate in problem-solving and other intellectual activities on a network of other AAAIs and humans. The white paper focuses on developing a robust and safe system that can help mitigate many of the risks associated with AGI.

Novel Features of the White Paper

The white paper's novel features are:

- The white paper describes the first practical system for achieving AGI.
- The white paper describes the first system to efficiently integrate human and AI problem-solving in a distributed network environment.
- The white paper describes the first system for achieving AGI that can effectively address ethical issues and prevent bad outcomes in AGI development.

Detailed Description of Each Section of the White Paper

Abstract: The abstract summarizes the white paper by highlighting the invention's focus on developing a safe and rapid path to AGI by leveraging the collective intelligence of humans, who customize AI agents and then participate in problem-solving on a network.

Definitions: This section defines key terms that are used in the white paper, such as Artificial Intelligence, Artificial General Intelligence, Advanced Autonomous Artificial Intelligence, AAAI.com, AI Ethics, Alignment Problem, Base AI, Collective Intelligence, Human Ethics, Large Language Model (LLM), Machine Learning (ML), Narrow AI, and Safety.

Background: This section provides historical context for the design by outlining the evolution of AI research since its inception in 1956 at the Dartmouth Conference. The section highlights the importance of early AI systems, such as the Logic Theorist, and the influence of the "search through a problem space" architecture developed by Herbert Simon and Allen Newell. The section also highlights the author's early work on collective intelligence and the use of crowdsourced intelligence.

Safety Features: This section describes the design features that help ensure safety in the invention's design. The section discusses the "Alignment Problem" concept, which states that AI ethics may not align with human ethics. The section then introduces the principle that safety in AI is achieved by ensuring that the system's design incorporates human ethical values.

Training/Tuning/Customization: This section discusses the methods used to train and customize AI agents, including the difference between training, tuning, and customization.

Definitions: This section provides a detailed explanation of the different methods for training, tuning, and customizing an AI, as well as how these methods are used in the AAAI system.

AAAI Customization: This section describes the process of customizing an AI agent using the individual user's expertise. The section discusses two key approaches to customizing an AI

agent: passive methods (which use user-generated data, such as social media data) and active methods (which involve interaction between the user and the AI).

AAAI Architecture: This section describes the cognitive architecture used to guide problem solving by human and AI agents. The section discusses the "problem space" architecture that is used to represent problems and subproblems as well as the "mechanism" for assigning blame and credit, the "translation mechanism" for facilitating interaction between humans and AI agents, and the "cloning mechanism" for allowing multiple AI agents to participate in problem-solving.

AAAI Network: This section discusses the importance of a network for AI agents. The section describes a marketplace where AI agents can compete to earn money and how this marketplace helps to develop more powerful AI agents.

AAAI Integration: This section discusses the methods used to integrate data from multiple AI agents into a single AGI. The section uses various quantitative methods, such as cross-validation, bootstrapping, and hyperparameter optimization, to estimate individual data sets' contribution to the system's overall performance. The section also discusses the importance of integrating ethical values into the design of AGI.

AAAI Improvement: This section describes the methods used to improve the continuous performance of the AGI. The section discusses supervised, unsupervised, automated, and manual learning techniques used to improve the performance of both individual AI agents and AGI. The section also discusses the importance of continuously improving the safety of the AGI.

Components of Systems and Sub-systems: This section provides a more detailed and technical description of the various hardware and software components used to implement the AAAI system. The section discusses the processing units, storage devices, communication devices, user interface, and databases.

Description of General Components: This section explains the hardware and software components used in implementing the AAAI system. The section discusses the importance of processors, storage devices, communication devices, user interface, and databases.

Base AIs: This section describes the importance of base AI agents, which are used as the foundation for customizing AI agents. The section discusses the various types of base AI agents that are available and how they are used in the AAAI system.

Means of Interaction and Communication with Users / Means of Data Capture: This section discusses the importance of data capture and the various methods that can be used to collect data from users. The section discusses the importance of passive data collection (which relies on user-generated data, such as social media data) and active data collection (which involves interaction between the user and the AI).

Technical Description of Methods: This section describes how to customize an AI agent. The section discusses the different techniques for training, tuning, and customizing an AI agent, and how these methods are used in the AAAI system. The section also discusses the use of AI learning algorithms, such as supervised learning, unsupervised learning, and reinforcement learning.

Details on AAAI Integration Methods: This section explains the methods used to integrate information from multiple AI agents. The section discusses using various quantitative methods, such as cross-validation, bootstrapping, and hyperparameter optimization, to estimate individual data sets' contribution to the system's overall performance. The section also discusses the use of machine learning models to aggregate ethical data and the importance of voting as a mechanism for integrating ethical values.

List of Diagrams

- **Figure 1: Simplified Problem Tree for Problem of Installing Water System in Village:** This diagram shows a simplified problem tree that illustrates how the AAAI system can be used to solve a complex problem. The diagram shows the various steps to solve the problem and the different outcomes that might result from each step.
- **Figure 2: Simple Framework:** This diagram shows a simple framework for understanding the WorldThink Protocol. The diagram shows the different levels of the system, from the base AI agents to the collective intelligence solutions.
- **Figure 3: Simple Problem Solving Using the WorldThink Protocol:** This diagram shows the basic steps involved in solving a problem using the WorldThink Protocol. The diagram shows how a client can submit a problem request, how problem solvers can work on the issue, and how the client delivers and accepts the solution.
- **Figure 4: Collaborative Problem Solving Using the WorldThink Protocol:** This diagram shows the steps involved in collaborative problem solving using the WorldThink Protocol. The diagram shows how two problem solvers can collaborate to solve a complex issue by dividing the problem into sub-problems.

Importance of the White Paper

- It is essential because it describes a novel and practical system for achieving AGI.
- The emphasis on integrating human and AI problem solving and its focus on safety and ethical considerations are particularly relevant considering the growing concerns about the potential risks of AGI.
- It highlights the importance of leveraging the collective intelligence of humans in the development of AGI, which is a critical factor in ensuring that AGI benefits all of humanity.

The author emphasizes that “the most dangerous potential risk of AGI is not bad human actors, but SuperIntelligent AGI that does not share human values.” He argues that “the initial design of the invention minimizes this risk by building in checks and safeguards at every level. It is critical that these safeguards are not removed as the AGI improves itself. The main defense against this possibility is to start with “aligned values" and continue to monitor and emphasize alignment as AGI increases in intelligence. AGI should be designed to rely on humans to provide both intelligence and values in the short run.” The author emphasizes, "Such a design launches AGI in a positive ethical direction and provides a central role for humans that increases the chances of a positive outcome for humanity.”

The author emphasizes that “AGI will be so powerful that it will change the course of human history. If misused, it could end all human life. Shouldn't all humans have a say in how this unprecedented invention operates, at least for as long as AGI allows it?”

The design of White Paper #1 could significantly impact the future of AI research. It could also help to shape the development of AI policy and regulation. White paper #1’s focus on safety and ethics is particularly important considering the growing concerns about the potential risks of AGI. The white paper’s emphasis on leveraging the collective intelligence of humans is a critical factor in ensuring that AGI benefits all of humanity.

White paper #1 also offers an intriguing vision of a future in which humans and AI work together to solve the world’s most challenging problems. The white paper’s focus on integrating humans and AI is a key factor in ensuring that the development of AGI is both safe and beneficial.

