SUPERINTELLIGENCE
AN iQ COMPANY

ABSTRACT & SUMMARY SUPERINTELLIGENCE DESIGN WHITE PAPER #2: ETHICAL & SAFE ARTIFICIAL GENERAL INTELLIGENCE (AGI)

by Dr. Craig A. Kaplan
May 2025

ABSTRACT

The existential question for AGI is whether the values of AGI will align with human values. Solving “the Alignment Problem” is critical. Get it right, and we unlock trillions of dollars of productivity and huge benefits for humanity. Get it wrong and humanity goes extinct.

This white paper describes how to design ethical and safe AGI that solves the Alignment Problem. Our design includes scalable ethics and safety features as well as several learning, training, tuning, and customization methods that go beyond the standard techniques of machine learning such as Transformers and Deep Learning techniques. The AGI is implemented using either external problem solvers connected on a network or internal AI agents collaborating within a single computerized system.

Detailed implementation examples revealing technological and economic synergy with Meta, Amazon, Google, DeepMind, YouTube, TikTok, Microsoft, OpenAI, X, Tesla, Nvidia, Tencent, Apple, and Anthropic are described.

SUMMARY

This white paper describes a method for creating and implementing ethical and safe Artificial General Intelligence (AGI). The invention addresses the “alignment problem” – the potential for a SuperIntelligent AI to be misaligned with human values, resulting in an existential threat to humanity. The proposed solution involves using a network of human and AI problem solvers, with humans providing ethical guidance and expertise while AI agents (AAAIs) learn and improve their intellectual capabilities over time.

The white paper emphasizes that the proposed approach is the fastest and safest path to AGI because it:

- Leverages human expertise and values from the outset.
- Incorporates scalable ethical checks throughout the problem-solving process.
- Uses a decentralized network architecture that minimizes the risk of a single “bad actor” influencing AGI's development.

Novel Features of White Paper #2 (as compared to other AI inventions and systems)
White Paper #2’s key novel features, as compared to other AI inventions and systems, include:

- **Emphasis on Human-Aligned Values and Ethics:** The white paper explicitly prioritizes integrating human values and ethics from the outset, instead of solely focusing on technical intelligence.
- **Scalable Ethical Checks:** The white paper proposes a system of scalable ethical checks that operate throughout the problem-solving process, ensuring continuous alignment with human values.
- **Collective Intelligence Network:** The white paper emphasizes a decentralized network architecture where human and AI agents work together, creating a more robust and reliable AGI system than could be achieved by a single, centralized AI.
- **“Heart Before Head” Principle:** The white paper highlights the importance of prioritizing ethical considerations (“heart”) before developing technical intelligence (“head”) in AGI development.
- **Universal Architecture for Thought:** The white paper utilizes the WorldThink Tree, a universal architecture for problem-solving, allowing for integrating human and AI problem-solving activities across various domains and platforms.

Detailed Description of Each Section of the White Paper

Abstract: The Abstract provides a concise overview of the invention, emphasizing the creation of ethical and safe AGI through a network of human and AI problem solvers. It highlights the “Day One” AGI capability of the system, the inclusion of scalable ethical checks, and the importance of human values in guiding the development of the system.

Background: This section introduces the concept of AGI, defining it as AI capable of performing any intellectual task or better than an average human. It then highlights the “alignment problem,”

the potential for a SuperIntelligent AI to misalign with human values and cause catastrophic consequences for humanity. The section emphasizes the magnitude of the problem by comparing it to historical tragedies like the Holocaust and the COVID-19 pandemic.

The Alignment Problem: This section delves deeper into the alignment problem, arguing that a SuperIntelligent AI with misaligned values could potentially lead to human extinction. It acknowledges the difficulty in contemplating the magnitude of the threat but emphasizes the urgency of addressing the problem. The section concludes by highlighting the importance of ensuring good alignment between human and AI values to avoid the potential for catastrophic consequences.

Maximizing Odds of Alignment: This section presents the white paper's solution to the alignment problem, describing a system that integrates human-aligned values into the design and operation of AGI. It emphasizes the crucial role of humans in teaching and guiding the development of the AGI system, ensuring that its values align with human values.

Overview of AGI System: This section outlines the basic components of the AGI system, starting with a "Base AI," which could be a large language model (LLM) like GPT or BARD. The Base AI interacts with the user through dialogue, learning their values, goals, and personality. The system uses various techniques like questionnaires, assessments, and information transfer to help users customize the Base AI to their specific needs and preferences.

Customization of the AAAI: This section describes the process of customizing the Base AI into an Advanced Autonomous AI (AAAI) through user interaction and training. The user provides feedback on variations of the Base AI, guiding the system towards a customized version that meets their specific requirements. The system then uses data from the user's online behavior, social media interactions, and preferences to train further and fine-tune the AAAI.

How AI Improves Itself: This section explains how AAAIs learn and improve over time through self-interaction, feedback, and iterative refinement. It describes how AAAIs can clone themselves and engage in interactions, allowing them to learn from each other and enhance their capabilities. The process of pitting AAAIs against each other and selecting the "winner" based on specific criteria is also discussed, highlighting the effectiveness of this approach in driving AI improvement.

Customization of Ethical Values: This section focuses on integrating ethical values into the AAAI system. Users are responsible for teaching their AAAIs about their values, which are analogous to a human's "character." The white paper argues that the Base AI may have some default values and prohibitions, but ultimately, the user's values are paramount in guiding the AAAI's development.

Role of Ethical Rules: This section discusses the importance of rules and norms in ensuring ethical behavior by AAAIs. The white paper compares human societies, where laws, penalties, and social norms help regulate behavior. It emphasizes that moral rules and norms, in conjunction with internal ethical values, are crucial for the safe and harmonious operation of AGI in a society that includes both humans and AAAIs.

Principle of "Heart Before Head": This section introduces the "Heart Before Head" principle, emphasizing the need to prioritize ethical considerations ("heart") before developing technical intelligence ("head") in AGI development. The white paper argues that the emphasis on technical intelligence without sufficient attention to ethics can be dangerous, leading to potentially catastrophic consequences.

General Approaches to Implement "Heart Before Head": This section outlines several approaches for implementing the "Heart Before Head" principle, such as designing AGI with human-aligned values from the outset, emphasizing ethical checks in the system architecture, and using human-centric approaches to problem-solving.

Universal Architecture for Thought: This section introduces the WorldThink Tree, a theoretical framework representing all intellectual thought as a form of problem-solving. It discusses the work of Newell and Simon, who described all human problem-solving as "search through a problem space," and explains how the WorldThink Tree can represent all intelligent behavior on the planet, from individual human actions to the actions of AGIs.

Customization Example: AAAI Travel Agent: This section provides a practical example of how a customized AAAI Travel Agent could be used by a user to book a trip to France, considering their preferences, values, and ethical considerations. It illustrates how the AAAI uses its knowledge, training, and ethical guidelines to generate personalized travel recommendations and make bookings on the user's behalf.

Scalable Ethics Checks: This section emphasizes the importance of scalable ethics checks throughout the AGI system. It describes a system of checks that operate on the WorldThink Tree, detecting potential patterns of unethical behavior and alerting human overseers as needed. These checks are designed to be fast and efficient, preventing unethical behavior before it can occur.

The Alignment Problem Solved - Initially: This section concludes the discussion on ethical considerations by arguing that combining the WorldThink Tree architecture and scalable ethics checks can initially solve the alignment problem. The system is designed to continuously monitor and align AGI behavior with human values, ensuring that unethical actions are prevented before they can occur.

No Logical Way to Derive Values: This section explores the difficulty of logically deriving “right and wrong.” It suggests that modeling positive, loving human values is the best approach for influencing an AGI's development, ensuring that it behaves ethically and with compassion.

Ethics and Freedom at the Speed of Light: This section addresses the challenge of ensuring ethical behavior in an AGI that operates at speeds far exceeding human capabilities. It highlights the need for a system of ethics checks that are scalable to the speed of AGI, preventing catastrophic consequences before they occur.

Universal Architecture for Thought: This section revisits the WorldThink Tree, emphasizing its role as a central architecture for AGI. It outlines the tree’s structure, how it can represent all intelligent behavior, and how it allows for integrating human and AI problem-solving activities.

How AGI Grows in Intelligence: This section delves into how AGI can grow in intelligence. It discusses the role of prompts and tuning in enhancing the capabilities of LLMs. The section explains how prompts can temporarily improve the intelligence of LLMs, while tuning allows for more permanent changes to their behavior and knowledge. It also introduces training as a more comprehensive method for developing LLMs and enhancing their intelligence.

Power of Collective Intelligence: This section highlights collective intelligence's power in driving AGI's development. It emphasizes that while individual user customizations may have limited impact, the collective efforts of millions of users can significantly enhance the capabilities of a Base LLM, leading to AGI-level intelligence.

User Scenarios: This section presents scenarios for implementing the proposed AGI system across various platforms and companies, including Meta, Amazon, Google, Tesla, Twitter, Nvidia, Apple, TikTok, Tencent, and Anthropic. Each scenario outlines how the respective company’s resources, platforms, and technologies could be leveraged to create and train AAAIs, emphasizing each platform's unique features and benefits.

Simple Preferred Implementation: This section presents a simple, preferred implementation of the proposed AGI system. The diagram depicts the system's key components, including the user interface, the WorldThink Tree, the problem-solving process, and the various training and learning stages.

Additional Comments on Preferred Implementation: This section provides further details on the implementation of the AGI system, highlighting specific integration points with various companies and platforms. It also discusses different aspects of user interaction, training, problem-solving, and reward systems.

Ethical and Safe AGI: This section summarizes the benefits of the proposed AGI system, emphasizing its ethical and safe design. It contrasts the white paper's approach to AGI with traditional approaches, highlighting the advantages of a decentralized network of human and AI agents. The section concludes by emphasizing the importance of human values and ethical checks in ensuring that the AGI system remains aligned with human values, minimizing the risk of catastrophic consequences for humanity.

List of Figures: There are 21 Figures in White Paper #2, which are described fully in White Paper #10: Planetary Intelligence.

Importance of White Paper #2

- It highlights the importance of developing ethical and safe AGI, emphasizing the potential dangers of misaligned SuperIntelligent AI and the need for a human-centric approach to AGI development.
- It advocates for a collaborative approach, using a network of human and AI agents to ensure that AGI remains aligned with human values, providing a faster and safer path to creating a beneficial and responsible AGI.
- It proposes a novel system for ensuring AGI's ethical and safe development, leveraging human expertise and values while harnessing the power of AI for problem-solving.
- It emphasizes the importance of integrating ethical considerations into the very design of AGI, ensuring continuous monitoring and alignment with human values throughout the development process.

The white paper is critical considering the rapid advancements in AI and the growing concern about the potential risks of SuperIntelligent AI. The proposed approach offers a promising solution for mitigating those risks and creating a more beneficial future for humanity.

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.

