SUPERINTELLIGENCE
AN iQ COMPANY

ABSTRACT & SUMMARY SUPERINTELLIGENCE DESIGN WHITE PAPER #4: Safe, Scalable, Artificial General Intelligence

by Dr. Craig A. Kaplan
May 2025

ABSTRACT

Artificial General Intelligence (AGI), when it arrives, will be the most powerful technology that has ever been invented.

Therefore, the ethical values that guide safe AGI must be democratic and broadly representative of the ethics and values of all of humanity. In contrast to existing approaches to AI safety, which rely on RLHF, constitutions, or ethical rules developed by a small set of engineers, this white paper describes systems and methods for obtaining a representative and statistically valid sample of ethical values from a wide range of humans.

The white paper also discloses novel methods for using and combining information from social media, knowledge modules, LLM weight matrices, and other sources. We describe new inventions designed to prevent hallucinations and errors by AI agents and to increase the audibility, transparency, reliability, scalability, and safety of AGI. Our design represents the fastest path to AGI because it builds upon and is synergistic with existing technology.

SUMMARY

This white paper describes a novel system for training Artificial General Intelligence (AGI) systems to be safe and scalable. It is based on Collective Intelligence (CI), where many individual AI agents are trained on a representative sample of human ethics and values and combined to align the resulting AI system with human values.

The design overcomes several limitations of existing approaches to AI safety, such as Reinforcement Learning with Human Feedback (RLHF) and Constitutional AI. RLHF is not scalable and struggles to adequately address the vast number of possible scenarios that might lead to unintended negative consequences. Constitutional AI relies on a set of ethical principles written by a small group of humans and may not reflect the values of humanity as a whole.

The white paper proposes a more scalable approach that relies on an extensive and diverse set of human-trained AI agents, each customized to represent a unique set of human ethical values. These AI agents are then combined to ensure that the resulting AI system is representative of the values of humanity as a whole.

The white paper also describes several methods for improving the efficiency and effectiveness of the training process, including:

- **Methods for combining weights from multiple AI agents.** The white paper describes several methods for integrating the weights of multiple AI agents, such as a simple linear combination. In this weighted combination, human input is given more weight than AI input, and a combination is based on the agents' expertise.

- **Methods for weighting input based on recency and other time-based factors.** The white paper describes several methods for weighting input based on time, including exponential decay, linear decay, and threshold-weighting.

- **Methods for dynamically flagging potential ethical issues in real-time.** The white paper describes a technique for dynamically flagging potential ethical problems in real-time and presenting these issues to other agents for resolution. This approach allows AI to learn from its mistakes and continuously improve its understanding of human ethics.

The white paper also provides several examples of how the design can be implemented, including a scenario where META uses its massive user base to create personalized AI agents aligned with individual users' ethical values.

Novel Features of the White Paper

The white paper proposes a novel approach to training safe, scalable, and aligned AGI by addressing several limitations of existing methods, such as RLHF and Constitutional AI.

The key novel features of the white paper include:

- **A Collective Intelligence (CI) approach** that relies on a large and diverse set of human agents, each of which has been customized to represent their own set of ethical values.

- **A method for combining the weights of multiple AI agents** in a way that ensures that the resulting AI system is representative of the values of humanity as a whole.

- **A method for weighting input based on recency and other time-based factors** ensures that AI systems are constantly updated with the latest ethical norms.

- **A method for dynamically flagging potential ethical issues in real-time** to allow AI to continuously learn from its mistakes and improve its understanding of human ethics.

- **A method for training AI to recognize and respond to dangerous scenarios**, which is essential for ensuring that AI systems are safe and ethical.

- **The use of knowledge modules**, which are essentially sets of training weights that can be combined with an AI system's existing weights to change its behavior in a known and predictable way.

- **A marketplace for knowledge modules**, where humans can share, trade, or license their knowledge modules.

The white paper proposes a new approach to AGI safety and alignment that addresses the limitations of existing methods and offers several novel features that could significantly advance the field of AI research.

Detailed Description of Each Section of the White Paper

Reference:
This white paper section references several previous papers on the current design. This includes Design White Papers #1 - #3:

- **White Paper #1: Advanced Autonomous Artificial Intelligence (AAAI)**
- **White Paper #2: System for Ethical and Safe Artificial General Intelligence (AGI)**
- **White Paper #3: System for Human-Centered AGI**

These previous white papers provide a foundation for the current design by describing the concepts of AAAI, Ethical and Safe AGI, and Human-Centered AGI.

Background: This section provides a general overview of AI and the development of AI agents. It describes the limitations of existing approaches to training AI agents and introduces the concept of Advanced Autonomous Artificial Intelligence (AAAI).

Problems with Current Approaches to AI and LLM Safety: This section discusses the limitations of existing approaches to AI safety, including Reinforcement Learning with Human Feedback (RLHF) and Constitutional AI. It argues that both approaches are not scalable and do not adequately address the challenges of ensuring that AI systems are safe and ethical.

Overview of the Design: This section provides an overview of the white paper's key innovations, including using a Collective Intelligence (CI) approach and developing a system for dynamically updating AI knowledge based on human values.

Description of Some Relevant Information Processing Systems: This section describes the general information processing systems used in the design. It explains that these systems can be implemented using a variety of hardware and software, including CPUs, GPUs, memory systems, and network communication systems.

Overcoming Problems with RLHF and Constitutional AI Safety Approaches: This section explains how the white paper's approach overcomes the limitations of RLHF and Constitutional AI by using a CI approach that relies on a large and diverse set of human agents. It argues that this approach is more scalable and representative of human values than existing approaches.

Contrasting Constitutional AI and the Current Design: This section contrasts the white paper's approach with Constitutional AI, arguing that the white paper's approach is more scalable, representative, and accurate than Constitutional AI.

Simple Implementations: Reinforcement Learning vs. Combining Weights: This section explains how the white paper's approach can be implemented using either a reinforcement learning or a weight combination approach. It argues that both methods are functionally equivalent and that the approach choice depends on the training process's specific circumstances.

Some Preferred Methods of Weight Combination: This section describes several preferred methods for combining weights from multiple AI agents, including a simple linear combination, a weighted combination where human input is given more weight than AI input, and a combination based on the agents' expertise.

Values or Ethics-Specific Implementation Considerations: This section discusses the challenges of training AI systems on ethics and values, including that there is no single correct answer to most ethical questions. It explains that the white paper's approach addresses these challenges using a representative sample of human values and dynamically updating AI's knowledge based on these values.

Ethical Solutions That Mirror What Humans Do: This section emphasizes that AI systems must be trained to behave in ways that mirror the ethical behavior of real humans. It argues that this can be achieved using a CI approach that relies on a large, diverse set of human agents.

Ethical Norms: This section discusses the importance of moral norms and how they can guide the development of ethical AI systems. It also provides examples of ethical standards commonly agreed upon by humans.

Ethical Contracts: This section discusses the importance of ethical contracts and how they can guide the development of ethical AI systems. It explains that humans often enter into ethical contracts when they join a group or participate in society.

The Safety Argument for Democratic, Representative Values: This section argues that a democratic and representative approach to training AI is more likely to lead to developing safe and ethical AI systems than other approaches, such as authoritarian or hierarchical approaches.

The Scientific Argument for Democratic, Representative Values: This section argues that a representative sample of human values is the most scientifically valid way to train AI systems on ethics and values. It explains that a representative sample is more likely to capture the true values of humanity as a whole than a smaller, more biased sample.

Efficient Training Methods: This section outlines the key constraints to be met when training safe and scalable AGI systems and describes a four-phase process for training AI systems that meets these constraints.

Detailed Implementation Example: This section provides a specific example of how a company like META can implement the white paper's design. It describes how META can use its massive user base and existing infrastructure to create personalized AI agents aligned with individual users' ethical values.

List of Diagrams
Diagrams are available in a separate file.

Importance of White Paper #4
This white paper is important because it proposes a novel and potentially groundbreaking approach to training safe, scalable, and aligned AGI.

The design has several advantages over the state of the art, including:

- It addresses the limitations of existing AI safety systems, such as RLHF and Constitutional AI, by proposing a more scalable, representative, and accurate approach to training AI.
- It addresses the challenge of ensuring that AI systems are aligned with human values using a CI approach that relies on a large and diverse set of human agents.
- It provides a framework for dynamically updating AI's knowledge based on human values, ensuring that AI systems are constantly updated with the latest ethical norms.
- It provides a detailed implementation example of how a company like META can use the design to create personalized AI agents aligned with individual users' ethical values.

The white paper's approach to AGI safety and alignment has the potential to significantly advance the field of AI research and make it possible to develop safe and beneficial AI systems that can be used to solve some of the world's most pressing problems.

Overall, the Design White Paper #4 proposes an important and innovative approach to the challenge of developing safe and scalable AGI that has the potential to significantly advance the field of AI research and make it possible to develop AI systems that are both safe and beneficial for humans.

