Serial #: PCT/US2024/017269
Title: **System and Methods for Safe Scalable Artificial General Intelligence**
Inventor: Craig A. Kaplan
Priority Date: February 28, 2023
Details: 174 total pages, 88 claims, 26 Figures

SHORT ABSTRACT

Artificial General Intelligence (AGI) is the most powerful technology ever 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 invention provides systems and methods for obtaining a representative and statistically valid sample of ethical values from a wide range of humans. This invention also discloses novel methods for using and combining information from social media, knowledge modules, LLM weight matrices, and other sources. The invention is designed to prevent hallucinations and errors by AI agents and increase AGI's audibility, transparency, reliability, scalability, and safety. The invention represents the fastest path to AGI because it builds upon and synergizes with existing technology.

GEMINI PRO SUMMARY
Provisional Patent Application #4 for Safe, Scalable, Artificial General Intelligence (AGI)

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

This approach is designed to overcome several limitations of existing Al safety systems, including Reinforcement Learning with Human Feedback (RLHF) and Constitutional Al. RLHF needs to be more scalable and adequately address the vast number of possible scenarios that might lead to unintended negative consequences. Constitutional Al 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 patent 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 Al system is representative of the values of humanity as a whole.

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

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

- **Methods for weighting input based on recency and other time-based factors.** The patent 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 patent describes a method for dynamically flagging potential ethical problems in real time and presenting these issues to other agents for resolution. This approach allows Al to learn from its mistakes and continuously improve its understanding of human ethics.

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

Novel Features of the Patent

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

The key novel features of the patent include:

- **A Collective Intelligence (CI) approach** relies on a large and diverse set of human agents, each customized to represent their own set of ethical values.
- **A method for combining the weights of multiple Al agents** in a way that ensures that the resulting Al 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 Al systems are constantly updated with the latest ethical norms.
- **A method for dynamically flagging potential ethical issues in real-time** to allow Al to continuously learn from its mistakes and improve its understanding of human ethics.
- **A method for training Al to recognize and respond to dangerous scenarios** is important for ensuring that Al systems are safe and ethical.
- **Knowledge modules** are sets of training weights that can be combined with an Al 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 patent 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 Al research.

Detailed Description of Each Section of the Patent

Prior PPAs Incorporated by Reference: This section of the patent incorporates several previously filed patent applications related to the current invention. This includes:

- **Advanced Autonomous Artificial Intelligence (AAAI) System and Methods**
- **System and Methods for Ethical and Safe Artificial General Intelligence (AGI), Including Scenarios with Technology from Meta, Amazon, Google, DeepMind, YouTube, TikTok, Microsoft, OpenAI, Twitter, Tesla, Nvidia, Tencent, Apple, and Anthropic**
- **System and Methods for Human-Centered AGI**

These earlier PPAs provide a foundation for the current invention by describing the concepts of AAAI, Ethical and Safe AGI, and Human-Centered AGI.

Background: This section provides a general overview of the field of Al and the development of Al agents. It describes the limitations of existing approaches to training Al 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 Al safety, including Reinforcement Learning with Human Feedback (RLHF) and Constitutional Al. It argues that both approaches must be more scalable and adequately address the challenges of ensuring that Al systems are safe and ethical.

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

Description of Some Relevant Information Processing Systems: This section describes the general information processing systems used in the invention. 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 patent's approach overcomes the limitations of RLHF and Constitutional Al 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 Invention: This section contrasts the patent's approach with Constitutional Al, arguing that the patent's approach is more scalable, representative, and accurate than Constitutional Al.

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

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

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

Ethical Solutions That Mirror What Humans Do: This section emphasizes that Al 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 ethical norms and how they can guide the development of ethical Al systems. It also provides examples of ethical norms commonly agreed upon by humans.

Ethical Contracts: This section discusses the importance of ethical contracts and how they can guide the development of ethical Al 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 Al is more likely to lead to developing safe and ethical Al 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 Al 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 that must be met when training safe and scalable AGI systems and describes a four-phase process for training Al systems that meet these constraints.

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

List of Diagrams
There are no diagrams in this Provisional Patent Application. Still, there are PCT and country applications based on the Provisional – contact iQ Company if diagrams are needed.

Importance of the Patent

This patent application is important because it proposes a novel and potentially groundbreaking approach to safe, scalable, and aligned AGI training. The invention is important for several reasons, including:

- It addresses the limitations of existing Al safety systems, such as RLHF and Constitutional Al, by proposing a more scalable, representative, and accurate approach to training Al.
- It addresses the challenge of ensuring Al 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 Al knowledge based on human values, ensuring Al systems are constantly updated with the latest ethical norms.
- It provides a detailed implementation example of how a company like META can use the invention to create personalized Al agents aligned with individual users' ethical values.

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

Overall, this patent proposes a critical and innovative approach to developing safe and scalable AGI that can significantly advance the field of Al research and make it possible to create Al systems that are both safe and beneficial for humans.