

Serial #: PCT/US2024/017261
Title: **Human-Centered AGI**
Inventor: Craig A. Kaplan
Priority Date: February 28, 2023
Details: 146 total pages, 74 claims, 18 Figures

SHORT ABSTRACT

Artificial General Intelligence (AGI) is safer if humans are kept “in the loop.” However, until now, no architecture for AGI existed that is both human-centered and highly scalable. The current invention of human-centered AGI includes humans in the loop and scales to super-human speeds while retaining human-aligned values. The invention contains novel systems and methods that include, without limitation: a) reputational methods that increase the efficiency and effectiveness of problem-solving by the (human and AI) intelligent entities that collaborate in the AGI system; b) use of LLMs’ abilities to understand and translate natural language into a universal problem-solving protocol; c) use of tree data structures combined with rewards to direct attention; and d) use of blockchain technology to reward problem solvers and capture a rigorous and auditable record of every cognitive step. Human-centered AGI can be implemented more rapidly than any existing approach.

GEMINI PRO SUMMARY

Provisional Patent Application #3 for Human-Centered Artificial General Intelligence (AGI)

This provisional patent application describes a novel system for creating safe and effective Artificial General Intelligence (AGI). The patent’s central thesis is that the fastest path to AGI is also the safest path. The invention is based on a human-centered network architecture that combines human and AI problem-solving agents, leveraging the strengths of both while mitigating the risks associated with purely AI-driven systems.

The patent proposes a network-based approach to AGI that leverages a typical human and AI cognition architecture. This architecture enables both humans and AI agents to collaborate in solving problems, harnessing their strengths while mitigating the risks of purely AI-driven systems. The network is based on the Human Problem Solving (HPS) framework, which provides a structured and rigorous method for problem-solving. The HPS framework is designed to be easily understood and used by both humans and AI agents, facilitating a collaborative and transparent process.

The patent emphasizes the importance of human values in the design and development of AGI. It advocates for a human-centered approach that ensures the AI system is aligned with human values and ethics from the outset. This is achieved through the integration of human problem-solving agents into the network and through the training of AI agents with human values. The patent acknowledges the potential risks of developing SuperIntelligent AI, particularly regarding alignment and safety. It argues that the proposed human-centered approach mitigates these risks by ensuring that the system has embedded human values and ethics.

The patent outlines several key benefits of the proposed system:

- **Safety:** The system mitigates the risks associated with misaligned or uncontrolled AI by integrating humans into the problem-solving process and training AI agents with human values.
- **Scalability:** The network architecture can be scaled to accommodate a wide range of human and AI solvers, enabling the system to handle complex and diverse problems.
- **Transparency and Auditability:** The system's rigorous and transparent architecture allows for the monitoring and auditing of the problem-solving process, increasing trust and accountability.
- **Speed:** The system's efficient and collaborative approach allows for faster problem-solving, enabling rapid progress toward the development of AGI.

The patent concludes by emphasizing the importance of the proposed invention, arguing that it represents the fastest and safest path to the development of AGI.

Novel Features of the Patent

The patent proposes a novel approach to the development of AGI that distinguishes itself from existing AI systems in several key ways:

- **Human-Centered Architecture:** The proposed system is based on a human-centered network architecture that incorporates both human and AI agents, leveraging their strengths while mitigating the risks associated with purely AI-driven systems.
- **Explicit Integration of Human Values:** The patent emphasizes the importance of human values in the design and development of AGI. It advocates for a human-centered approach that ensures the AI system is aligned with human values and ethics from the outset.
- **Transparency and Auditability:** The system's architecture is designed to be transparent and auditable, allowing for the monitoring and auditing of the problem-solving process and increasing trust and accountability.
- **Incremental Learning:** The system leverages incremental learning, enabling AI agents to learn from the experience of humans and other AI agents, leading to continuous improvement.
- **Democratization of Ethical Decision-Making:** The patent proposes a democratic approach to ethical decision-making, allowing human owners of AI agents to train their agents with their own values and ethics.

Detailed Description of Each Section of the Patent

The patent is organized into several sections, each addressing a key aspect of the proposed system:

- **Abstract:** The Abstract outlines the patent's central thesis: the fastest path to AGI is also the safest path. The invention is based on a human-centered network architecture that combines human and AI problem-solving agents, leveraging the strengths of both while mitigating the risks associated with purely AI-driven systems.
- **Background:** This section provides context for the patent by outlining the current state of AI research and the challenges associated with developing AGI. It highlights the importance of solving the alignment problem to ensure that AGI is aligned with human values and ethics.
- **System and Methods for Human-Centered AGI:** This section describes the core system architecture and methods for implementing the human-centered AGI system. It highlights the use of a typical architecture for human and AI cognition, the importance of human values in the system's design, and the role of a rigorous problem-solving framework.
- **How to Build AGI in the Fastest, Safest Manner:** This section describes the process of building the human-centered AGI network. It emphasizes the importance of a network-based approach that leverages the collective intelligence of human and AI problem-solving agents.
- **Benefits of a Common Architecture for AI and Human Cognition:** This section outlines the benefits of the proposed common architecture for human and AI cognition, including the ability to avoid unintentional errors, enable automatic learning, ensure scalability and modularity, and maximize safety.
- **Avoiding Unintentional Errors:** This section explains how the common architecture helps to avoid unintentional errors by providing a rigorous framework for problem-solving that ensures that AI agents are aligned with human values and ethics.
- **Enabling Automatic Learning:** This section describes how the common architecture enables automatic learning by allowing AI agents to learn from the experience of both humans and other AI agents.
- **Enabling Scalability:** This section explains how the common architecture enables scalability by allowing the system to accommodate a wide range of human and AI solvers.
- **Enabling Modularity and Scalability:** This section explains how the common architecture enables modularity and scalability by allowing new AI agents to be easily integrated into the system.

- **Maximizing Safety:** This section discusses the importance of safety in the development of AGI and highlights how the proposed system's architecture and human-centered design maximize safety by ensuring that the AI system is aligned with human values and ethics.
- **One Preferred Implementation of the AGI Network:** This section describes a specific implementation of the human-centered AGI network based on the Human Problem Solving (HPS) framework.
- **The Theory of Human Problem Solving:** This section provides background information on the HPS framework, a model of human problem-solving developed by Newell and Simon.
- **Why HPS Works for AI Agents:** This section explains how the HPS framework can be applied to developing AI agents, enabling them to learn and improve their problem-solving capabilities.
- **Easy for Humans to Participate:** This section describes how humans can easily participate in the HPS-based problem-solving process, contributing their expertise and knowledge.
- **Required Systems and Methods for AGI Network Already Exist:** This section explains how the necessary systems and methods for implementing the human-centered AGI network already exist in computer programming.
- **AGI Network Solves the “Representation Problem”:** This section discusses the importance of problem representation in AI and how the proposed AGI network addresses this challenge by combining the strengths of human and AI problem-solving agents.
- **Multi-Modal Representations:** This section explains how using multi-modal representations, incorporating visual, auditory, and other sensory data, can enhance the intelligence of AI agents.
- **LLMs Facilitate Human-AI Interaction:** This section discusses the role of large language models (LLMs) in facilitating communication between humans and AI agents.
- **HPS Highly Scalable:** This section explains how the HPS framework is highly scalable, enabling the system to handle a wide range of problems.
- **The Role of Attention:** This section discusses the importance of attention in problem-solving and how the AGI network addresses this challenge by guiding human and AI solvers toward the most relevant problems.
- **Learning via Proceduralization of Knowledge (Solutions):** This section discusses how AI agents can learn and improve their problem-solving capabilities by learning from the experience of other AI agents and humans.
- **Unique Approach to AGI:** This section argues that the proposed approach to AGI is unique and innovative, differentiating it from other AI research and development efforts.

- **Human Training of AAAs Influences Safety:** This section emphasizes the importance of training AI agents with human values to ensure they behave ethically.
- **Democratization of Ethical Values in Safety:** This section discusses the importance of democratizing ethical decision-making in AI by empowering human owners of AI agents to train their agents with their own values.
- **Role of System Rules and Norms in Safety:** This section explains how the AGI network can further incorporate rules and norms to enhance safety and ethical behavior.
- **Role of Reputation in Safety:** This section discusses how the AGI network can leverage reputation and track records to promote ethical behavior and foster user trust.
- **Safety Checks at the Speed of AI Thought:** This section explains how safety checks can be integrated into the AGI network to ensure that ethical considerations are considered during problem-solving.
- **Implementation Example:** This section provides a specific implementation example of the human-centered AGI network, illustrating how the system can train and deploy AI agents.

List of Diagrams

The provisional patent includes no diagrams added later in the PCT and country applications.

Importance of the Patent

This patent is significant for several reasons.

First, it proposes a novel and innovative approach to the development of AGI that addresses the critical challenge of alignment and safety. The patent's human-centered approach emphasizes the importance of incorporating human values and ethics into the system's design, offering a potentially safer alternative to purely AI-driven systems.

Second, the patent is forward-looking, recognizing the importance of AI-driven problem-solving in the context of rapid technological advancements and the increasing role of AI in various aspects of human life. It anticipates the need for a scalable, flexible, and adaptable system that can evolve and adapt to changing needs and circumstances.

Third, the patent highlights the importance of collaboration between human and AI agents, recognizing their strengths. It proposes a system that leverages the combined intelligence of both humans and AI to solve complex problems, ultimately leading to a more efficient and effective problem-solving process.

Finally, the patent provides a detailed and comprehensive overview of the system architecture and methods for implementing the proposed human-centered AGI system. This level of detail is crucial for facilitating the implementation and testing of the system by researchers and developers, potentially leading to significant advancements in AI research and development.

The patent's proposed human-centered approach to AGI is a significant departure from traditional AI development paradigms, which often focus on developing ever-more powerful AI systems without adequately addressing the challenges of alignment and safety. By emphasizing human values, ethics, and collaboration in its design, this patent presents a compelling vision for the development of safe and effective AGI that can benefit humanity.