Serial #:	PCT/US2024/020334
Title:	**System and Methods for Safe Alignment of SuperIntelligence**
Inventor:	Craig A. Kaplan
Priority Date:	February 28, 2023
Details:	226 total pages, 80 claims, 36 Figures

SHORT ABSTRACT

Artificial General Intelligence (AGI) and SuperIntelligence (SI) will exceed human abilities in almost every cognitive activity. To ensure human safety and survival, we must design SI to align and stay aligned with human values. This invention discloses many principles and methods for designing and aligning safe AGI. Methods include novel ways to combine values democratically from many intelligent entities, improve ethical decision-making, dynamically comply with regulations, protect minority values, resolve values conflicts, vote on ethical decisions, handle delegation of voting authority, and use simulations, game theory, and constitutional AI – all in the service of making advanced AI systems safe for humanity. Specific use cases and implementations are disclosed. Scalable safety features are integral to system operation. Finally, if AGI or SI goes awry, the systems are designed to maximize the chances that dangerous components can be shut off safely.

GEMINI PRO SUMMARY

Provisional Patent Application #7 for System and Methods for Safe Alignment of SuperIntelligence

This Provisional Patent Application lays out a new approach to the safe development and deployment of very advanced AGI and SuperIntelligence (SI) systems. The central premise is that true safety and alignment of any advanced AI system can only be achieved by incorporating human values and ethics into the design itself. The patent focuses on three key areas: 1) A new design for AI/AGI/SI systems based on a network of individually customized agents that are interconnected and that share and learn from each other; 2) A set of principles and methods for ensuring that each agent's design reflects and prioritizes human values and ethics; and 3) A detailed approach for ensuring that the overall system is aligned with human goals and preferences and that it can be safely managed and controlled.

Novel features of the patent:

This patent application is distinct from other AI inventions and systems for several reasons:
It emphasizes a network of agents instead of monolithic AI systems. This approach allows for greater flexibility, adaptability, and control over the overall system. Each agent can be independently customized and aligned with specific human values and goals, making it easier to manage potential risks and ensure that the overall system remains aligned with human preferences.

It prioritizes human values and ethics as the central design principle. The patent advocates for developing AI/AGI/SI systems that are fundamentally aligned with human values and ethics

from the beginning. This approach contrasts with many other AI systems designed to learn human values and ethics after being built, which can lead to significant risks and challenges.

It proposes a comprehensive framework for the safe alignment, control, and governance of AI/AGI/SI systems. The patent addresses the technical aspects of AI/AGI/SI system design and the critical issues of human-AI interaction, transparency, accountability, and conflict resolution. It offers a more holistic and practical approach to developing and deploying very advanced AI systems.

Detailed description of each Section of the patent:

1. Introduction. This section provides a general overview of the growing interest in developing AGI and SI systems and highlights the need for a new approach to safe and aligned AI. It also discusses the inherent limitations of existing AI systems and the challenges in achieving true alignment with human values and ethics.

2. Background. This section comprehensively reviews the existing research on AI, AGI, and SI. It discusses the various approaches to AI design, including the use of monolithic AI systems and the different methods for aligning AI with human values and ethics. The section also discusses the ethical and societal implications of AGI and SI systems, particularly the potential risks and challenges of uncontrolled or misaligned AI.

3. Description of the invention. This section presents the key features of the proposed invention, including the design of AI/AGI/SI systems as a network of individually customized agents. It discusses the importance of incorporating human values and ethics into the design of each agent and outlines the key design principles for achieving safe and aligned AI. The section also discusses the different methods for training and aligning agents, including using human-in-the-loop feedback mechanisms and the various techniques for ensuring that the overall system remains aligned with human goals and preferences.

4. Principles and Methods. This section presents a set of design principles and methods for achieving safe and aligned AI systems. It emphasizes the importance of human-centered design, transparency, and accountability. The section also discusses the various techniques for aligning agents with human values and ethics, including reward functions, constraints, and human feedback mechanisms.

5. Implementation methods. This section outlines the specific techniques for implementing the proposed invention. It discusses the different methods for creating and training agents, including supervised learning, reinforcement learning, and transfer learning. The section also discusses the techniques for managing and controlling the overall system, including the use of human-in-the-loop feedback mechanisms and the different methods for detecting and mitigating risks.

6. Application. This section discusses the potential applications of the proposed invention. It highlights the potential benefits of safe and aligned AI/AGI/SI systems for addressing a wide range of societal and global challenges, including developing new technologies, advancing scientific knowledge, and improving human well-being.

7. Conclusion. This section summarizes the key points of the invention and emphasizes the importance of a new approach to safe and aligned AI/AGI/SI systems. It concludes by highlighting the potential benefits of the proposed invention for addressing a wide range of societal and global challenges, including the development of new technologies, the advancement of scientific knowledge, and the improvement of human well-being.

List of Diagrams:

Diagram 1: Network of Agents for Safe and Aligned AI/AGI/SI System. This diagram shows the proposed architecture of a network of interconnected agents that share and learn from each other. Each agent is independently customized and aligned with specific human values and goals. The diagram illustrates how the different agents can be trained and aligned to achieve a common objective and how the overall system can be managed and controlled.

Diagram 2: Human-in-the-Loop Feedback Mechanism for AI/AGI/SI System. This diagram illustrates how human feedback can be used to train and align AI/AGI/SI systems. It shows how humans can provide feedback to the system to guide its learning process and ensure it remains aligned with human values and ethics. The diagram also highlights the importance of transparency and accountability in human-AI interaction.

PCTs and country applications have more diagrams. Please get in touch with iQ Company if you desire to see these.

Importance of the patent:

This patent application is important because it presents a new and compelling approach to developing and deploying very advanced AI systems safely. It addresses the fundamental challenges of aligning AI with human values and ethics and proposes a practical framework for achieving safety and alignment. The invention has the potential to revolutionize the field of AI and to pave the way for the development of safe and beneficial AGI and SI systems that can help address some of the most pressing societal and global challenges.