SUPERINTELLIGENCE
AN iQ COMPANY

ABSTRACT & SUMMARY SUPERINTELLIGENCE DESIGN WHITE PAPER #6: CATALYSTS FOR GROWTH OF SUPERINTELLIGENCE

by Dr. Craig A. Kaplan
May 2025

ABSTRACT

Data is the “fuel” that powers the machine learning “engine” for Artificial Intelligence. However, identifying high-quality data that can catalyze smarter AI, AGI, and SuperIntelligent systems is becoming an increasingly challenging bottleneck for machine learning. This whitepaper not only describes novel methods for identifying the most valuable data, but it also presents an entirely new framework for understanding the information content of AI-relevant datasets. The methods can be used by intelligent systems autonomously or in collaboration with humans.   Novel methods for accelerating AI learning and updating the knowledge of AI systems in real time are also disclosed. Consistent with the view that human survival may depend on the fastest path to AGI, also being the safest path, the white paper describes catalysts that help maximize alignment between the values of AGI and humans. These innovative catalysts increase not only the intelligence but also the safety of AI systems.

SUMMARY

White Paper #6 describes a novel approach to developing safe and ethical Artificial General Intelligence (AGI) and SuperIntelligent AI systems. It emphasizes the importance of collective intelligence, a network of human and AI agents, instead of relying on a single, monolithic LLM.

White Paper #6 focuses on three key aspects of AI systems:

1. **Information Acquisition:** The white paper proposes new methods for identifying and acquiring relevant and useful information for increasing AI systems' intelligence. This includes expanding the traditional Shannon-sense information theory to incorporate “differences” as a measure of information, rather than simply relying on probabilities and surprise.
2. **Representation:** The white paper highlights the importance of using high-level representations for problem-solving instead of relying solely on low-level representations

like bits or tokens. It suggests that adopting such representations can significantly accelerate the development of intelligent systems.

3. **Safety:** The white paper emphasizes the importance of ensuring safety and ethical considerations in designing and developing AI systems. It proposes using a combination of human oversight, automated simulation methods, and adversarial testing techniques to achieve alignment between human and AI values, thus reducing the risks associated with the uncontrolled growth of SuperIntelligent AI.

Novel Features of the White Paper

White Paper #6 features several novel aspects that distinguish its approach from other AI approaches:

1. **Collective Intelligence:** The white paper emphasizes the need for a collective intelligence network of human and AI agents, as opposed to a single, monolithic LLM.
2. **Kaplan Information Theory (KIT):** The white paper proposes a new theory of information, KIT, which goes beyond the traditional Shannon-sense information theory by incorporating "differences" as a key measure of information.
3. **Goal-Relatedness:** The white paper introduces the concept of goal-relatedness as a crucial dimension of information, emphasizing its importance in guiding the acquisition of relevant information.
4. **Human-AI Collaboration:** The white paper proposes a collaborative approach to AI development that integrates human insights and values into the design and training of AI systems.
5. **Automated Safety Mechanisms:** The white paper describes automated mechanisms, such as adversarial testing and simulation methods, to ensure that the development of AI systems prioritizes safety and ethics.
6. **High-Level Representations:** The white paper emphasizes the importance of using high-level representations for problem-solving instead of relying solely on low-level representations.
7. **Continuous Learning and Optimization:** The white paper proposes a continuous learning and optimization loop for AI systems that leverages feedback from both human users and automated simulations and synthetic data.
8. **Delegation Strategy:** The white paper emphasizes the importance of developing a delegation strategy that balances human oversight with the increasing capabilities of AI systems.

9. **Community of Agents:** The white paper suggests that a community of PSIs – each reflecting the values of a human owner – can collectively minimize the risks associated with a single, malevolent SI.

Detailed Description of Each Section of the White Paper

The white paper is divided into several sections, each focusing on a specific design aspect. Here is a detailed summary of each section:

1.0 Overview of the Design:

- This section provides a general overview of the white paper's purpose and scope. It emphasizes developing safe and ethical AGI and SuperIntelligent AI systems.

2.0 Previous White Papers

- This section lists the prior Design White Papers #1 - #5 upon which the current design builds. These white papers cover previous work by the author on the development of safe and ethical AI and SuperIntelligence systems.

3.0 Background for the Design:

- This section provides a brief background on the history and theoretical underpinnings of AI, including a discussion of the contributions of key AI pioneers such as Marvin Minsky, Claude Shannon, Allen Newell, and Herbert Simon.

3.1 Introduction:

- This section highlights the significance of the Dartmouth Conference in 1956, which is considered the birthplace of the field of Artificial Intelligence. It also discusses the contributions, framed as “gifts”, of four key AI pioneers: Marvin Minsky, Claude Shannon, Allen Newell, and Herbert Simon.

3.2 Gift #1: Society of Mind (Minsky):

- This section discusses Minsky's concept of a "Society of Mind," where intelligence emerges from the collective behavior of many smaller processes called "agents."

3.3 Gift #2: Information Theory (Shannon):

- This section reviews Shannon's contributions to information theory, emphasizing the importance of Shannon's Entropy as a measure of information. It also discusses the limitations of Classical Information Theory.

3.4 Gift #3: Problem-Solving Theory (Newell and Simon):

- This section describes Newell and Simon's "Problem Solving Theory," which provides a framework for representing and solving problems, including the importance of "operators" and "goal-relatedness."

3.5 Gift #4: Bounded Rationality (Simon):

- This section discusses Simon's concept of "bounded rationality," emphasizing that limited information processing capabilities constrain human behavior.

3.6 Conclusion of Background:

- This section summarizes the key insights from the background information provided in the previous sections, leading to a vision of a safe and ethical SuperIntelligent AI system that incorporates the best aspects of the work of AI pioneers.

3.7 References:

- This section lists relevant references cited in White Paper #6.

3.8 Additional Contextual Information for This Design:

- This section provides additional contextual information related to the design's importance, including the need for a collective intelligence approach to AGI, the potential for Planetary Intelligence, and the significance of ensuring the alignment of AI values with human values.

4.0 Classical Information Theory Contrasted to Kaplan Information Theory:

- This section contrasts Classical Information Theory (as proposed by Shannon) with Kaplan Information Theory (KIT), highlighting the limitations of Classical Information Theory and introducing the core concepts of KIT.

4.1 Inventive Methods as Understood by Classical Information Theory:

- This section describes how the inventive methods described in the white paper can be understood within the framework of Classical Information Theory.

4.2 Limitations of Classical Information Theory:

- This section outlines the limitations of Classical Information Theory in terms of its inability to fully account for the relative nature of information, the importance of “news,” and the role of context.

4.3 Kaplan Information Theory (KIT):

- This section provides a detailed explanation of KIT, emphasizing the importance of "differences" as a measure of information.

4.4 Multiple Dimensions of Information in KIT:

- This section outlines the multiple dimensions of information considered in KIT, including the importance of surprise, goal-relatedness, knowledge bases, and the cost of acquiring information.

4.5 Estimating the Value of Information:

- This section discusses estimating the value of information using KIT, taking into account multiple dimensions of difference.

5.0 Inventive Methods:

- This section introduces inventive methods for acquiring and using information efficiently, including a basic process for identifying, acquiring, and ingesting information.

5.1 Methods Relevant to Classical Information Theoretical Notions of Information as Entropy:

- This section focuses on methods relevant to the classical definition of information related to entropy and rarity, including using Kolmogorov complexity and compression algorithms.

5.1a Kolmogorov Complexity and Compression for Determining Information Content:

- This section further explores the use of Kolmogorov complexity and compression algorithms for determining the information content of datasets.

5.1b Cross Entropy and KL Divergence:

- This section discusses the use of cross-entropy and KL Divergence as measures of information and their potential applications in AI systems.

5.1c Limitations of Entropy-Related Methods:

- This section outlines the limitations of entropy-related methods in capturing the full scope of information, including the importance of goal-relatedness and context.

5.2 Goal-relatedness Methods:

- This section introduces the concept of goal-relatedness as a crucial dimension of information, emphasizing its importance in guiding the acquisition of relevant information.

5.3 Mathematical Specification of Relevance:

- This section describes a mathematical framework for quantifying the relevance of information, considering goal-relatedness and Shannon Entropy.

5.4 A Simple Evaluation Function for Seeking Useful Information:

- This section proposes a simple evaluation function for prioritizing the acquisition of helpful information, based on goal-relatedness, relevant knowledge, information content, and cost.

5.5 Innovative Methods for Estimating Kaplan Information:

- This section discusses innovative methods for estimating Kaplan information, including the importance of goal-relatedness and the use of human feedback in the evaluation process.

5.5a Importance of Representation:

- This section emphasizes the importance of using appropriate representations for AI systems.

5.5b One Method for Estimating Information Value & Catalyzing Intelligence Growth:

- This section proposes a specific method for estimating the value of information and catalyzing intelligence growth.

5.6 Automated Methods and Safety Considerations:

- This section discusses automated methods for knowledge acquisition and the importance of safety and ethical considerations in designing and implementing AI systems.

5.6a Automated Simulation Methods:

- This section describes automated simulation methods for testing and validating AI behavior.

5.6b Realtime Scenario Creation Methods:

- This section discusses real-time scenario creation methods to test and validate AI behavior.

5.6c Adversarial Testing Methods:

- This section introduces the concept of adversarial testing as a method for ensuring the safety and ethical use of AI systems.

5.6d Simultaneous Scenarios:

- This section discusses the use of simultaneous scenarios to explore the potential for AI systems to "jailbreak" themselves and reveal vulnerabilities.

6.0 Inventive Catalysts for Increasing Intelligence Beyond Information Seeking:

- This section introduces a new set of inventive catalysts for increasing intelligence beyond information seeking, focusing on the importance of high-level representations, the acquisition of new representations, and methods for accelerating intelligence.

6.1 Importance of High-Level Representations:

- This section discusses the importance of using high-level representations for problem-solving and the limitations of relying solely on low-level representations.

6.2 Acquisition of New Representations:

- This section discusses the importance of acquiring new representations for AI systems and provides examples of how to do so.

6.3 KIT-based Heuristics and Methods to Accelerate Intelligence:

- This section describes heuristics and methods that leverage KIT to accelerate the growth of intelligence.

6.3a Catalyzing Effects of Higher-Level Representations:

- This section discusses the effects of using higher-level representations on the speed and efficiency of intelligence growth.

6.4 Methods for Assessing Artificial Intelligence:

- This section discusses the challenges of measuring AI intelligence and proposes using standardized human intelligence tests, crowdsourcing, and non-standardized creative problem-solving tasks.

6.4a Extension of Standardized Tests of Human Intelligence to AI:

- This section proposes using standardized tests of human intelligence to measure AI intelligence.

6.4b Crowdsourcing Evaluation of AI Intelligence:

- This section describes the use of crowdsourcing to evaluate the quality of AI solutions.

6.4c Use of Non-Standardized Creative Problem Solving / Insight Tasks:

- This section suggests using non-standardized creative problem-solving tasks to assess the cognitive abilities of AI systems.

6.5 Methods to Modify (Optimize) the Personality of PSI:

- This section discusses methods for modifying the personality of a PSI to improve its interaction style with humans and other AI systems.

6.6 Methods for Scalable Delegation as Intelligence Increases Exponentially:

- This section proposes a delegation strategy that balances human oversight with the increasing capabilities of AI systems.

6.7 Safety via a Community of Agents Approach to AGI:

- This section discusses the benefits of a community approach to AGI development, emphasizing the importance of collective intelligence and consensus-building.

7.0 One Preferred Implementation of Some Methods in an AI/AGI/SI/PSI System:

- This section provides a specific example of how the inventive methods can be implemented in a real-world AI/AGI/SI/PSI system.

8.0 Concluding Remarks:

- This section summarizes the key points of the design, emphasizes the importance of safe and ethical AI development, and concludes with a call to action for AI researchers.

List of Figures: There are 24 figures in White Paper #6, which are described fully in White Paper #10, Planetary Intelligence.

Importance of White Paper #6

1. **Novelty and Utility:** The white paper introduces novel approaches to AI development, including a new theory of information, a focus on goal-relatedness, and a collaborative approach to human-AI interaction. These approaches are highly relevant to the current challenges and opportunities in AI research.
2. **Addressing the Alignment Problem:** The white paper provides a framework for addressing the Alignment Problem, which is ensuring that AI systems align with human values and ethics. This is a critical issue for the safe and ethical development of SuperIntelligent AI.
3. **Scalability:** The white paper proposes scalable methods for developing and deploying AI systems, emphasizing the importance of delegation and the use of collective intelligence.

4. **Human-Centered Approach:** The white paper emphasizes a human-centered approach to AI development, recognizing the importance of human values, ethics, and oversight in the design and development of AI systems.

5. **Safety First:** The white paper emphasizes that safety and ethics must be incorporated into the design and development of AI systems from the outset.

6. **A Roadmap for the Future of AI:** The white paper provides a roadmap for the future of AI, suggesting a path toward developing safe, ethical, and powerful AI systems.

White Paper #6 significantly contributes to AI research by proposing novel approaches to developing safe, ethical, and powerful AI systems. The design's focus on collective intelligence, goal-relatedness, human-AI collaboration, and continuous learning and optimization is highly relevant to AI researchers' current challenges. The white paper provides a framework for developing AI systems that are both powerful and safe, ensuring that the development of SuperIntelligent AI is aligned with human values and ethics.

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.