**Ethics Essay: Artificial Intelligence: A Guide for Thinking Humans**

**Melaine Mitchell**

**Chitradevi Maruthavanan**

**PSU ID: 950828319**

Melanie Mitchell's book "Artificial Intelligence: A Guide for Thinking Humans" takes readers on a personal journey to understand artificial intelligence, its limitations, and how it differs from human intellect. The book presents this potentially dry topic as a fascinating story with cultural references, making it thought-provoking and informative. The book covers a wide range of topics including the history and origins of AI, its pros and cons, symbolic and non-symbolic ways, and their limitations. It also discusses the examples and limitations of convolutional and deep neural networks and makes predictions for the future. The author focuses on the history and breakthroughs of AI and its transformation, describing cutting-edge AI programs and modern AI and machine learning models. The book is easy to read and enlightening, discussing not only how AI works but also what it means to be intelligent in the human sense.

McCarthy came up with the term "artificial intelligence" to differentiate it from cybernetics. There have been developments in AI with two approaches: symbolic and sub-symbolic. Symbolic AI programs use words or phrases that humans can understand as knowledge and rules to combine and process symbols to complete tasks. An early AI program was called the general problem solver or GPS. The symbolic approach to AI argues that mimicking the brain with algorithms isn't necessary to achieve computer intelligence. Instead, they believe that a symbol-processing algorithm can capture general intelligence completely.

Sub-symbolic AI is the other significant trend in AI. Symbolic AI was initially influenced by mathematical reasoning and the way people expressed their conscious thought processes. In contrast, sub-symbolic AI was inspired by neuroscience and aimed to capture the often-unconscious cognitive processes that drive fast perception, such as recognizing faces or understanding spoken phrases. Sub-symbolic AI systems use a stack of equations with operations on numbers, rather than human-readable language. The efficiency and performance of these networks rely greatly on hyperparameters like learning rate and the number of hidden units and layers. Finding the best set of hyperparameters is a challenging task. While reinforcement learning is a popular approach to teaching AI how to play games, there are concerns about how algorithms learn. Q-Learning is a popular technique for implementing reinforcement learning. It rewards good behavior (winning and good choices) and punishes bad behavior (losing and bad choices).

AI went into a period of stagnation after Minsky and Papert discredited the perceptron. This event became an early example of the cyclical nature of the AI industry, with a repeating pattern of bubbles and crashes. The cycle consists of two phases. During Phase 1, new ideas generate optimism within the research community, with the news media often exaggerating the potential breakthroughs in AI. Academic research and commercial start-ups receive funding from government financiers and venture investors. Phase 2 follows, where the promised breakthroughs fail to materialize or fall short of expectations, leading to a decrease in government and venture capital investment. As a result, AI research slows down, and start-up companies often fail. The AI community now recognizes this pattern as the "AI spring," followed by overpromising and media hype, and then the "AI winter."

Humans use language to communicate, and AGI (Artificial General Intelligence) aims to enable computers to understand human language. Natural Language Processing (NLP) is the study of combining languages and computers, which has made significant progress in addressing issues such as language translation and sentiment analysis. This progress has been achieved using statistical models and various data representation methods. However, a significant concern is that current AI systems do not truly understand or sense language in the same way that humans do, despite the game AlphaGo astonishing the world's best players. Leading businesses have invested heavily in AI research and development, but the author believes that AI has a low chance of breaking through the meaning barrier. While state-of-the-art AI systems have almost matched or outperformed humans on some tasks, they are unable to understand the subtle nuances in human perception, language, and thought. Although AI has made significant progress in various fields, each development remains limited to its respective narrow field.

The objective is to create an "Artificial General Intelligence" (AGI) that can understand and sense the world like humans, as well as comprehend the outcomes of its actions. Once achieved, humanity will no longer be the most intelligent species. It's uncertain whether AGI will benefit the world, and based on progress and patterns, the singularity is predicted to occur in several decades. The book focuses on the journey towards achieving "human-level" intelligence in AI and its effects on future AI developments. The author explores the contrast between AI's hype and actual accomplishments, providing personal insights and perspectives in an engaging manner.