Emily Sheetz
Artificial Intelligence
Literature Survey

Universal Artificial Intelligence

I. Introduction

As artificial intelligent agents progress and become capable of solving increasingly complex problems, the possibility of general machine intelligence becomes more real. Shane Legg investigates many concepts of general artificial intelligence in his doctoral thesis, “Machine Super Intelligence” (2008). He works from an informal definition of general intelligence to a formal mathematical definition that can be used to evaluate intelligence in biological and artificial systems. A number of researchers have addressed the concepts related to developing general artificial intelligence. Though the results of the research preceding and following Legg’s work are largely theoretical—and in some cases speculative—they are formalized mathematically and founded on principles sound enough to allow for analysis. By examining the literature related to Legg’s dissertation on general machine intelligence, a survey can be created about the significant topics in this area of artificial intelligence research.

II. Definitions of Intelligence

Before defining universal artificial intelligence, Legg explores the complexity of the definition of intelligence itself. The definition of human intelligence, though intuitive, is difficult to articulate concretely. The validity of tests that measure human intelligence are often questioned because of the possibility of bias by race, gender, class, or culture and their inability to test all of the different facets of intelligence. When considering artificial intelligence, the definitions become even less clear and it becomes apparent that current intelligence tests are anthropocentric, using human intelligence as the benchmark. Legg points out that “as technology advances, our concept of intelligence will continue to evolve with it.” Rather than definitively explain what intelligence is and how it can be measured, he acknowledges the inherent complexity of discussing intelligence in both biological and artificial systems.

After examining a number of definitions of intelligence and acknowledging their commonalities and differences, Legg does provide an informal definition of intelligence upon which he bases his dissertation: “Intelligence measures an agent’s ability to achieve goals in a wide range of environments.” One of the key concepts commonly associated with intelligence is the ability to learn and adapt in order to achieve some goal. Legg’s formalization of universal artificial intelligence utilizes this notion as well.

Universal artificial intelligence involves creating intelligent agents that behave optimally in a variety of unknown environments. This definition places no restrictions on the internal workings of the agent or on the time efficiency of the agent. Legg points out that such agents are still largely theoretical, as many discussions of universal artificial intelligence assume infinite computational resources, which in practice is not the case. However, some promising research towards developing general agents is described in a later section of this survey (see Section V).

III. Formalization of Intelligent Agents

Universal artificial intelligence relies on the ability of the agent to learn about its environment and learn to improve its actions from its mistakes. In his dissertation, Legg outlines the agent-environment framework. Based on observations or perceptions from the environment, the agent can choose actions that maximize a reward, which indicates to the agent how beneficial its current situation is in the context of achieving its predetermined goal. The agent-environment framework is central to reinforcement learning. Through reinforcement learning, agents can learn to modify behavior according to the rewards or penalties it receives from the environment.

Throughout the literature, a distinction is made between passive and active environments. In passive environments, general agents exhibit passive inductive learning, which involves drawing inferences from observations of the environment in order to make predictions or complete sequences. While these are useful abilities, the agent has no effect on the environment. However, in active environments, agents can take actions to affect the future states of the environment. The actions the agent takes are based off of its predictions derived from observations. Oftentimes agents are goal driven in active environments.

In both passive and active environments, the key abilities of the intelligent agent involve making predictions and learning. Inductive inference and reinforcement learning allow agents to carry out these tasks.

Through inductive inference, agents can determine causes of events given the observations of the environment. From these inferences, predictions can be made about future states of the environment. Effective inductive inference depends on Occam's razor, which states that the simplest hypothesis that supports the evidence history is the most likely candidate for a good model of the environment. In the case of agents in active environments, this means that the actions the agent takes are rational if they make the most sense given the previous observations. In an erratic environment, this does not necessarily mean that the agent will be the most successful; however, it is still considered rational. Inductive inference also draws on probability theory, specifically Bayes' Rule, which computes the conditional probability of an event given previous observations. Bayes' Rule allows the belief of a certain events to be updated based on new evidence.

Through inductive inference and reinforcement learning—as formalized in terms of probability theory and the agent-environment framework by Legg in his dissertation—a universally intelligent agent can theoretically make predictions, interact with, and learn about how to maximize rewards in a wide variety of environments.

IV. Measuring and Testing Intelligence

With the possibility that machine intelligence could surpass human intelligence, researchers are investigating ways to measure the intelligence of an agent. These tests would be general enough to evaluate both biological and artificial agents ranging in intelligence from specialized to universal. Intelligence measures would lead to the creation of an intelligence order relation that would compare the intelligence of all such agents. Because current artificial intelligent systems are powerful in typically one particular domain, these systems with specialized intelligence would be ordered lower in the relation than systems with general intelligence. Without the ability to draw a comparison to powerful general agents, it is only possible to speculate as to where human intelligence would be ordered by the intelligence measure (Legg, 2008).

A number of tests have been proposed since the beginnings of research in artificial intelligence. In his dissertation, Legg briefly examines the benefits or drawbacks of various intelligence tests. Several such tests are outlined in this section, based on related research that more formally derives each test.

Turing Test

The Turing Test, also known as the imitation game, determines whether or not an agent can convince human judges that it is human. If the agent accurately mimics human behavior and the judges cannot discriminate its behavior from that of a computer, then the agent is said to be intelligent. This test requires the system to possess a significant understanding of human behavior and interactions (Legg, 2008).

The Turing Test is criticized for being a pass-fail test that gives preference to agents with intelligence that mimics human intelligence. Rather than being a true test of intelligence, the Turing Test is considered to be a test of humanness (Legg, 2005). Machines could theoretically pass this test without possessing any "real intelligence" by using large look-up tables to store answers to questions that imitate answers the human judges would expect (Legg, 2007).

In response to these criticisms, derivatives of the Turing Test attempt to change the objective of the test by imitating multiple facets of intelligence. These tests involve developing intelligent agents that can pass for a human toddler or for small mammals (Legg, 2007).

C-Test

Legg's dissertation briefly touches on the C-Test for machine intelligence. The derivation of the C-Test, or comprehension test, is formally outlined in the paper by Hernández-Orallo, published in 2000, which is cited by the seminal work. The following discussion of the C-Test is based on this paper.

The C-Test is proposed as a non-anthropocentric test defined in computational terms that does away with the Turing Test by evaluating a variety of cognitive abilities, but focusing on comprehension ability. Hernández-Orallo defines comprehension as the ability to understand by observing evidence and constructing a plausible model. The C-Test assesses comprehension through a series of prediction tests, in which the agent needs to predict the next symbol in a sequence. The test favors the simplest "correct" answer. Each test that contributes to the agent's score is weighted based on the difficulty of the questions and the time it took for the agent to come up with a solution.

The validity of the C-Test for measuring intelligence in both biological and artificial systems is supported by the fact that the results of the C-Test correlate with IQ test scores for humans. Additionally, it distinguishes between knowledge and liquid intelligence, or the ability to solve problems. However, the validty C-Test is called into question in that it simulates a static, passive environment, which does not require the agent to interact with or affect its environment (Legg, 2005; Legg, 2007).

Universal Intelligence

The universal intelligence measure is discussed in depth in Legg's dissertation. The beginning work on formalizing this measure was presented in a paper by Legg and his dissertation supervisor Hutter in 2005, which is cited by the dissertation. Information on the universal intelligence measure will be drawn primarily from this work, which laid the groundwork for some of the big concepts presented in the seminal work of this survey.

Unlike the static C-Test, the universal intelligence measure is an interactive test because it requires the agent to interact with its environment. This measure is also dynamic because, in order to increase its intelligence score, the agent must learn from and adapt its behavior (Legg, 2008).

The agent is represented by a function, π, which inputs the history of observations, rewards, and actions and outputs a new action for the agent to take based on the conditional probability of the action over the current history. The environment is represented by a function, μ, which computes the conditional probability of the next observation and reward sequence given the current history of observations, rewards, and actions.

The universal intelligence measure is represented by the equation,

$$\Upsilon(\pi) := \sum_{\mu \in \boldsymbol{E}} 2^{-K(\mu)} V_{\mu}^{\pi}$$

where π is the agent function, μ is the environment function, and **E** is the set of all possible environments. Legg's dissertation explains the purpose of this equation in detail, but the overview provided by Legg and Hutter in their 2005 paper is sufficient. To understand the function conceptually, we should know that K(μ) represents the complexity of the environment and V_{μ}^{π} represents the weighted future rewards. The complexity term is designed to favor simple solutions—in agreement with Occam's razor—with minimal program lengths and computation time. The weighted rewards term is designed to favor agents that seek to maximize rewards into the future rather than in the immediate present.

The universal intelligence measure outputs a real value that is independent of how other systems have scored. As a result, scores can be compared among agents that have been evaluated using this measure. Such a comparison is called an Intelligence Order Relation (IOR), which could evaluate systems that range from unintelligent random agents to super intelligent agents. Furthermore, this measure is free from many of the criticisms that tests of human intelligence receive because it is free from bias (Legg 2008).

The downside of the universal intelligence measure is that the part of the equation that deals with the complexity of the environment is not computable and very difficult to approximate. Additionally, the measure is the sum of the agent's performance over *all* possible environments, which is impractical to compute. A potential solution is to sum the performance over a reasonably sized representative sample of environments. Approximations of complexity and samples of environments could allow the universal intelligence measure to be approximated.

Anytime Intelligence Tests

One of the criticisms of the measure of universal artificial intelligence presented in Legg's dissertation is that computation time is not considered. In fact, in the definition of universal artificial intelligence upon which the work is based, efficiency is explicitly not considered, though it is acknowledged to be an important concern in practical contexts. In an attempt to redefine intelligence tests to take time into consideration, Hernández-Orallo and Dowe define the anytime intelligence tests in their 2010 paper, which cites the seminal work.

The purpose of the anytime tests is to develop a series of tests that can accurately evaluate systems of any intelligence level that operate on any time scale. The researchers consider a number of possible tests and examine the benefits and drawbacks of each in evaluating intelligence. These tests generalize the C-Test from strictly passive environments to tests that include active environments as well. The key differences between the anytime tests and the universal intelligence measure is that the anytime tests sample environments rather than take the sum of performance over all possible environments, require that the environments be reward-sensitive to the actions the agent takes, progressively alter the complexity of the test environments, average rewards by the number of actions, and take time into account—either as the time since the last action taken or as the time taken relative to the complexity of the test environment.

The suite of anytime intelligence tests can be combined to test for different aspects of intelligence. For example, one test could be used to evaluate the speed with which the agent makes decisions, while another could evaluate the intelligence of the agent without respect to time.

Support for the anytime intelligence tests comes when an equivalent, more familiar test is considered. The anytime intelligence tests are compared to collections of videogames in which players play for a short amount of time and the difficulty of each game is adjusted as the player plays more games. This connection is promising, especially when considering the Arcade Learning Environment, which is the next intelligence test to be discussed. However, as the paper formalizes these tests mathematically but does not experimentally implement these tests, it can only be concluded that these tests would provide a rough approximation to the intelligence of an agent.

Arcade Learning Environment

The Arcade Learning Environment (ALE) is presented in a paper written by Bellemare, et al. in 2013 that cites the seminal work. The ALE is a platform for evaluating general artificial intelligence that provides an interface to hundreds of Atari 2600 games. The variety of games require the agent to interact with and learn from diverse and complex environments.

The paper formalizes how the ALE could be used for developing general agents. A number of games would be used as training games, while other unseen games would be used as testing games. General agents trained on these games would have to be able to effectively plan their strategies—the game spaces are too large to allow for exhaustive searches—and to learn from the environments through reinforcement learning—rewards, or scoring points, in each game is sometimes difficult to predict due to sparse distribution of rewards and difficulties of extracting relevant state information from game screens.

In order for the ALE to be an effective measure of general intelligence, an agent’s scores would have to be normalized across all games and then aggregated together into a single intelligence score.

Though it was not experimentally tested, it seems to be promising as a test for general intelligence because it reflects a wide range of complex problems. Furthermore, the research presented in the paper would generalize to other game playing systems, should general intelligent agents master playing Atari 2600 games. Other playing systems with better graphics and even more complicated games could serve as new tests for general intelligence.

Discussion of General Intelligence Tests

Having explored the possible benefits and drawbacks of a number of different intelligence tests, let us discuss implications of applying such tests to both biological and artificial systems. Legg's and Hutter's 2007 paper not only defines machine intelligence, but elaborates on the qualities of valid intelligence tests. Based on the common criticisms addressed in both the 2007 paper and Legg's dissertation, a discussion about the significance of comparing human and artificial intelligence can be formed.

One of the common criticisms of many of the proposed general intelligence tests—in particular, the universal intelligence measure—is that equations do not address different facets of intelligence, such as creativity or imagination, nor do they address consciousness. Legg and Hutter address that because the human brain is likely just a "meat machine," it is unlikely that human machinery is the optimal machinery for intelligence. However, an argument in favor of the significance of consciousness could be developed if the evaluation of the universal intelligence measure—or any intelligence test—showed an upper bound for machines and showed that humans perform above this limit. This would prove that humans must possess some unique element to the operation of their intelligence that causes a significant difference when compared to machine intelligence. However, without any direct comparison between human intelligence and universal artificial intelligence, these arguments are purely speculative.

V. General Agents

Because the concepts related to universal artificial intelligence are largely theoretical and frequently assume that agents have infinite computational resources, completely general agents have not yet been implemented. However, by scaling down the ideas behind general intelligence, attempts can be made at developing universally intelligent agents.

Legg dedicates a portion of his dissertation to discussing Marcus Hutter's work in formalizing a general agent which behaves optimally in a wide range of interactive environments, which he calls the AIXI agent. After taking a look at Hutter's theoretical general agent and the work that led up to the mathematical formalization, we can look at how attempts at general agents have been derived based on the theory behind the AIXI agent.

AIξ Model and AIXI Agent

In Hutter's 2000 paper, he begins to hypothesize about the AIξ model, which lays the groundwork for the more formal AIXI agent, which is discussed in Legg's dissertation. Hutter's paper will be used to discuss the theoretical concepts behind the AIξ model and the powerful AIXI agents.

The AIXI agent is a universal intelligent agent that learns about its environment through interactions and modifies its behavior to optimize performance based on the rewards it receives. The formal definition of the AIXI agent utilizes the AIξ model, which is unique in that it converges to behavior that matches that of the optimal agent designed for a particular environment. The AIξ model is incredibly powerful in that its generality still allows it to utilize many concepts of artificial intelligence, such as probability theory, utility theory, reasoning, reinforcement learning, and knowledge engineering and representation.

However, the AIξ model is incomputable because, as most theoretical general intelligent agents do, it assumes unbounded computational resources. In the best case, the AIξ model can be approximated. A computable variation is the AIξ^{tl} model, in which time t and space l bounded. While bounding the time and space available to the agent that implements the AIξ^{tl} model allows the model to be computable, the computation time is still quite large, making the model suitable for small "toy environments." An effective implementation of the AIξ model would need to scale the problem down through greater restrictions on the environment space ξ.

Legg and Hutter further elaborate on definitions of machine intelligence and the universal intelligence measure in their 2007 paper. Their papers from both 2005—from which we based or explanation of the universal intelligence measure—and 2007 are drawn from heavily in Legg's dissertation. The more in depth 2007 paper explains that, according to the universal intelligence measure discussed in Section IV, the AIXI agent is highest on the intelligence order relation and, as a result, is provably optimal. Though the AIξ^{tl} model is incomputable and as a result has practical complications, the AIXI agent has theoretically reduced the problem of general agents to computational questions. While these computational questions are difficult to solve, the reduction in the problem of general intelligence is significant.

MC-AIXI

A Monte-Carlo AIXI approximation, abbreviated MC-AIXI, is presented as a preliminary implementation of the AIXI agent in the 2011 paper by Veness, et al., which cites Legg's dissertation. There was doubt for some time whether the theoretically optimal AIXI agent would motivate the development of practical algorithms. However, by scaling the problem down, the MC-AIXI implements many of the mathematical concepts of the AIXI agent.

Veness, et al. review that the AIXI agent seeks take actions that maximize rewards and use inductive inference to make predictions about future states of the environment based on experience. In order to approximate the AIXI agent, a general agent must be able to plan its future actions and learn from the environment. The developed general agent is able to plan its future actions using a derivative of the Monte-Carlo tree search and is able to learn using context tree weighting, which allows the agent to process new experiences efficiently. Using these techniques, the authors developed the MC-AIXI agent, which approximates the mathematical solution to general intelligence presented by the AIXI agent.

The MC-AIXI agent was tested experimentally in a number of problems, such as mazes with different rewards schemes, Tic-Tac-Toe against a random opponent, Rock-Paper-Scissors against an opponent with a predictable bias in its strategy, and Pacman, which was partially observable with regards to the locations of ghosts and food pellets when not in the immediate vicinity. In all of these, the MC-AIXI agent's performance converged towards optimal performance as it gained more experience. In the incredibly challenging domain of the partially observable Pacman game, the agent did not as closely approximate optimal play, but learned important concepts, such as not running into walls, seeking food, and running away from ghosts. The agent did not learn to chase ghosts when it ate a power pill and sometimes acted erratically when not in the vicinity of food or ghosts. However, the MC-AIXI agent is promising in the fact that it did learn to perform effectively in such a large and challenging environment. These results demonstrate the possibility of scaling MC-AIXI up to even larger problems in order to more closely approximate general intelligence.

Limitations of General Agents

Having examined theoretical and experimental attempts at formalizing a general agent, it is important to examine the limitations of universal artificial intelligence. In his dissertation, Legg points out that there will never exist a general agent that is able to perform optimally in all environments. Though general agents will converge to optimal performance in a wide variety of environments given that it learns from sufficient experience, it is unrealistic for a general agent to match the performance of the optimal agent for every environment. The overall goal of universally intelligent agents must be to perform close to optimally in as many environments as possible. Such an agent, though not optimal in *every* environment, would be powerful enough to learn to solve a significant number of problems without any previous knowledge of many diverse environments.

VI. Implications of Universal Artificial Intelligence

Having formalized the mathematics behind universal artificial intelligence and explored many of the theoretical concepts and practical limitations of general agents, it is important to consider the implications of success in this area of research. In the last section of the seminal paper, Legg discusses some of the philosophical questions related to achieving general machine intelligence. Chambers analyzes the philosophical implications of universal artificial intelligence in his 2010 paper that cites Legg's dissertation. Chambers' work considers the possibility of not only universal artificial intelligence, but super intelligence, in which machines are able to create new generations of machines more intelligent than themselves, creating an intelligence explosion culminating in singularity.

Further advancements in universal artificial intelligence would likely be accompanied by both an intelligence explosion and a speed explosion, in which the intelligence and speed of machines progress towards infinity. These explosions could lead to a number of possible consequences, including beneficial ones such as “a cure for all known diseases, an end to poverty, extraordinary scientific advances,” or dangerous ones such as “an end to the human race, an arms race of warring machines, the power to destroy the planet.”

Chambers argues that singularity is a real possibility through a series of logical arguments. One of his arguments states that because evolution produced human intelligence, similarly, humans will contribute to the evolution of artificial intelligence. This idea is particularly interesting considering that artificial evolution is proposed as a possible method for developing universal general intelligence (Legg, 2008). Another of Chambers’ arguments is that if artificial intelligence is produced using methods that can be abstracted and extended, then these methods will, in fact, be extended to create super intelligent agents. Based on these and a variety of other arguments, singularity is presented as not unlikely in the future.

However, Chambers acknowledges that there could be obstacles to the singularity. Such obstacles could include limits in intelligence space, which claims that we are already near an upper limit of intelligence and could not progress further; human intelligence being a poor starting point for creating super intelligent systems; a lack of motivation to develop super intelligence; or a motivation against developing super intelligent machines.

If singularity is a possibility, then it may be helpful for humans to consider how to intervene in the development super intelligence. One possible constraint on universal artificial intelligence would be to limit the cognitive capacities of agents so that they are helpful to humans but cannot obtain key features like autonomy. Another is to give these agents specific values, such as a value for human survival or obeying human commands. These values could be introduced through direct programming or through learning. Trying to give intelligent agents values through learning releases much of human control over the end result, and only allows us to affect the evolutionary algorithm or process. However, even if the desired values were learned by the super intelligent agents, there is no guarantee that these values would be passed on to the next generation of agents.

Chambers ends his paper by considering the possible roles humanity will play in a world in which singularity is achieved. He proposes four possible options: "extinction, isolation, inferiority, or integration." Ultimately, if humanity is to survive and maintain a shred of its identity and consciousness—which throughout the literature has been noted as a key difference between human and artificial intelligence (Legg, 2008)—then humans would have to integrate through a process of uploading and enhancing human experience with the singularity.

The entirety of Chambers' paper, while based on credible sources on universal artificial intelligence—one of which is the seminal work of this survey, Legg's dissertation on machine super intelligence—is speculative. However, the philosophical questions he poses have been asked throughout the growth of the field of artificial intelligence. Furthermore, as consistent with several of the works considered in this survey, the paper emphasizes the importance of discussing the philosophical questions raised by developments in artificial intelligence.

In his dissertation, when Legg considers the implications of the development of universal artificial intelligence, and possibly even super intelligence, he notes that "The defining characteristic of our species is intelligence. […] If our intelligence were to be significantly surpassed, it is difficult to imagine what the consequences might be" (Legg, 2008). Even if we cannot predict how human intelligence will be challenged by the development of super intelligence nor how human intelligence would compare to universal artificial intelligence according to an intelligence ordering relation, the consequences of these developments would be immense.

VII. Conclusion

Shane Legg’s doctoral thesis, “Machine Super Intelligence,” lays the groundwork for the development and testing of universal artificial intelligence. He draws on previous research done in collaboration with Hutter that led directly to the work presented in his dissertation and several works that outline possible tests for evaluating general intelligence in biological and artificial systems. The research presented in his paper leads to new formalizations for general intelligence tests and relates to attempts at implementing the impractical but provably optimal mathematical solutions to general agents by scaling environment spaces and potential problems down to more manageable constraints. Legg’s work also continues the conversation about the consequences of progressing the field of artificial intelligence and creating agents whose universal intelligence could rival that of humans. By examining how the concepts related to universal artificial intelligence are developed and discussed throughout the literature related to Legg’s “Machine Super Intelligence,” we can obtain an understanding of the significant issues, current challenges, successful attempts at implementation, possibilities of future research, and philosophical consequences of progress in this area of artificial intelligence research.

References

Bellemare, Marc G., et al. “The arcade learning environment: An evaluation platform for general agents.” *Journal of Artificial Intelligence Research* 47 (2013): 253-279.

Chalmers, David. “The singularity: A philosophical analysis.” *Journal of Consciousness Studies* 17.9-10 (2010): 7-65.

Hernández-Orallo, José, and David L. Dowe. “Measuring universal intelligence: Towards an anytime intelligence test.” *Artificial Intelligence* 174.18 (2010): 1508-1539.

Hernandez-Orallo, Jose. “Beyond the Turing Test.” *Journal of Logic, Language and Information* 9.4 (2000): 447-466.

Hutter, Marcus. “A theory of universal artificial intelligence based on algorithmic complexity.” *Arxiv* (2000).

Legg, Shane, and Marcus Hutter. “A universal measure of intelligence for artificial agents.” *International Joint Conference on Artificial Intelligence* 19 (2005): 1509-1510.

Legg, Shane, and Marcus Hutter. “Universal intelligence: A definition of machine intelligence.” *Minds and Machines* 17.4 (2007): 391-444.

Legg, Shane. *Machine super intelligence.* Diss. University of Lugano, 2008.

Veness, Joel, et al. “A Monte-Carlo AIXI Approximation.” *Journal of Artificial Intelligence Research* 40.1 (2011): 95-142.