Ethical Considerations of the Approaching Technological Singularity

Jeff Wildgen
CSE 5290 – Artificial Intelligence
Summer 2011

***Abstract*—The prevailing consensus in the subject of machine ethics is that humanity is grossly unprepared for the theoretical creation of an artificial super-intelligence. History provides countless examples of scenarios where a superior technology, either natural or artificial, is able to gain an evolutionary advantage over an inferior. The warnings are becoming so much more apparent that our complacency in this topic may no longer be acceptable. Our future of a species may well come to depend on the policies we enact today.**

I. Introduction

The concept of creating mechanical devices to automate laborious or mundane tasks can be traced back throughout most of recorded human history. As described in Homer's Iliad, the ancient Greeks envisioned a race of mechanical servants built by the god Hephaestus [Levene Gera, 2003]. More recently, it was discovered that Leonardo da Vinci developed detailed conceptual drawings of a mechanical knight based on his human anatomical research. Analysis of the drawings indicates that the device would have been able to sit up, wave its arms, and move it's head [Taddei 2007]. At the arrival of early 20th century, the Czech writer Karel Čapek coins the term "robot" while presenting the concept of an artificial device that is now widely regarded as the modern version of an android [Čapek, 1920].

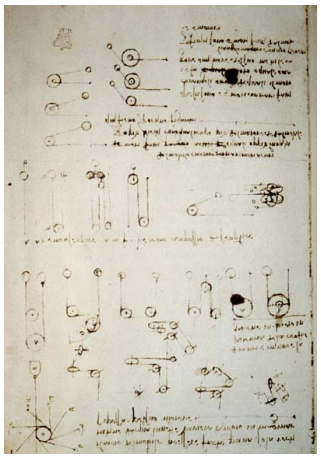

Figure 1: Drawing of Leonardo da Vinci's Robot

With the beginnings of the modern era, mathematics and computer science have given rise to a new form of automation that exists in a purely logical sense. Analogous to the mechanical robots that had been conceived throughout time, these new software based "bots" have a very similar role, but exist completely within the abstract realm of information and logic. These automata are now beginning to possess the abilities to replicate a myriad of human functions including speech recognition, visual pattern matching, and fundamental logical reasoning. In addition to these capabilities, the current pace of research is driving toward an ever-increasing level of autonomous behavior. Self-preservation, replication, and growth; basic elements of human life; are increasingly emerging as favorable traits for incorporation into our creations.

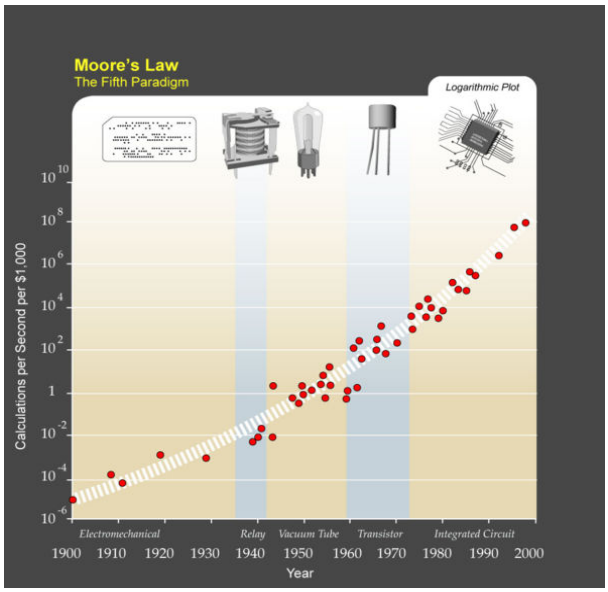

Figure 2: Moore's Law (R. Kurzweil)

As physical hardware computing power and algorithmic sophistication continuously progress, it is becoming more and more apparent that human society will need to confront the possibility that one of these creations will reach a recognizable level of self-awareness. Given the fantastic pace of technological achievement in the previous century, one must assume that over the course of this century, humans will inevitably succeed in building a machine that is capable of generating and hosting a mind such as ours [Hibbard 2008]. Once this event occurs, our society will bear witness to the moment the evolution of information organization reaches its most significant turning point. Nature, in its timeless drive to bring order to chaos, will have created a society of beings with such mastery of information that they are able to manipulate it toward

the creation of a new artificial being. Bill Joy, in his essay, "Why the Future Doesn't Need Us", provides a compelling argument in favor of how the evolution of machine cognition coupled with our own inability to keep pace will drive our species to extinction [Joy, 2003]. This is strong motivation for humanity to at least begin considering how we as a species can remain relevant in the rapidly approaching future.

II. THE INEVITABLE SINGULARITY

Rapid change in technology is a fact our society faces every day. To illustrate this, consider the actor James Cagney who was born in 1899. During his lifetime of 86 years, he witnessed the pace of human technology change from the horse and buggy, to the automobile, to powered flight, atomic energy, space travel, and the beginnings of personal computers. The natural implication of this is that at some point, our computing power coupled with our understanding of cognitive science, will reach a point where we create a system that can rival our own capacity to create new knowledge. Technology ethicists have coined the term "Singularity" to denote this event.

The Law of Accelerating Returns

Ray Kurzweil, in his book "The Age of Spiritual Machines" puts forth the concept of "The Law of Accelerating Returns". In this capacity, he is able to describe how the pace of technological advancement moves as an acceleration curve instead of a linear one. Kurzweil asserts that humans live in a linear world, but technology progresses in an exponential one. The doubling has already become apparent, as we have witnessed processing power double approximately 32 times since the end of the Second World War. This means that a year's worth of advancement today could very well equate to an hour's worth of advancement in the future [Kurzweil 1999].

The key factor behind this concept is that a human being cognitively processes his world from the perspective of a relatively short moment. Humans are therefore so well adapted to managing change that they become immersed in this moment and tend to not notice the accelerating pace of advancement. Only through the perspective of history can they see how quickly the changes of the past year have been relative to those of the past decade [Kurzweil 2001]. It is because of this point, we as a society need to be able to raise our level of awareness and recognize that the pace of technological change will likely force us to deal with the ramifications of our creations before we are able to actually understand them fully.

Seed AI and Recursive Enhancement

One of the emerging ideas that may directly contribute to the arrival of the singularity is a theory called Seed AI. This concept describes a type of AI that is initially programmed with a minimal level of intelligence, but also with the ability to recursively enhance itself [Yudkowsky, 2001]. In the course text, Russell and Norvig mention that one of the most significant dangers may end up being a technology that grows faster than we are able to control it [Russell 2010]. While the concept is not new to science fiction, the theories behind Seed AI show that this technology might be possible in reality. Recursive growth in machine intelligence would likely progress slowly and well within the means of our capacity to control it at first. However the very real danger arises when it exponentially rises at a pace faster than our ability to outsmart it. Bostrom describes this type of device using the term "Superintelligence" [Bostrom, 1998]. This type of intelligence would be vastly more capable than our own for solving problems. This event is also marked as the point where humanity will have reached the singularity. Given the current level of complacency, it will likely be that we as a society are completely unprepared.

Because of the danger this technology poses to humanity's very existence, it is imperative that we consider an evaluation of our own ethical standards. Society's scientists and engineers have the duty and moral obligation to ensure their research pursuits are

Kurzweil's *Law of Accelerating Returns*

- Evolution applies positive feedback in that the more capable methods resulting from one stage of evolutionary progress are used to create the next stage. As a result, of progress of an evolutionary process increases the rate exponentially over time. Over time, the "order" of the information embedded in the evolutionary process (i.e., the measure of how well the information fits a purpose, which in evolution is survival) increases.
- A correlate of the above observation is that the "returns" of an evolutionary process (e.g., the speed, cost-effectiveness, or overall "power" of a process) increase exponentially over time.
- In another positive feedback loop, as a particular evolutionary process (e.g., computation) becomes more effective (e.g., cost effective), greater resources are deployed toward the further progress of that process. This results in a second level of exponential growth (i.e., the rate of exponential growth itself grows exponentially).
- Biological evolution is one such evolutionary process.
- Technological evolution is another such evolutionary process. Indeed, the emergence of the first technology creating species resulted in the new evolutionary process of technology. Therefore, technological evolution is an outgrowth of–and a continuation of–biological evolution.
- A specific paradigm (a method or approach to solving a problem, e.g., shrinking transistors on an integrated circuit as an approach to making more powerful computers) provides exponential growth until the method exhausts its potential. When this happens, a paradigm shift (i.e., a fundamental change in the approach) occurs, which enables exponential growth to continue.

conducted in a safe and controlled manner. No system is completely foolproof, but we must be sure we are carefully understanding and weighing the risks.

III. DESIGNING WITH SAFEGUARDS

There are several ideas that address our ability as a society to prepare for the potential creation of an artificial intelligence that could one day surpass our own. This isn't a new concept however. We as a society already possess the capability to destroy ourselves, and our world many times over. It could even be argued, given the factors of scale and historical context, that this has been a problem we have been living with for a very long time. Therefore, it is conceivable to entertain the notion that we already know how to deal with this problem. We seem to have been able to bring ourselves to the brink of destruction, but suddenly conjure enough rationalism to stop short. This lends toward some optimism that we will be able to also assemble a strategy to coexist with or even complement a superhuman intelligence brought on by a post-singularity world. However, as Bill Joy has pointed out, if we don't attempt to place or embed rules into our technology, we will lose the opportunity to have a say in how they will govern themselves [Joy, 2003].

The Three Laws of Robotics

One of these concepts deals with the subject of embedding safeguards into the very heart of the technology as it is created. An early proponent of this concept was the science fiction writer, Isaac Asimov. In his 1942 short story, "Runaround", he first introduced the concept of the "Three Laws of Robotics", and devoted most of his writing career to exploring the ramifications of these Laws. Encoded in each of these laws is a fundamental property that no machine should ever be designed in such a way as it could cause deliberate harm to humanity. Each of Asimov's laws has been restated below.

Asimov's *Three Laws of Robotics*

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey any orders given to it by human beings, except where such orders would conflict with the First Law
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law

The important factor behind this is that it is one of first times we began to think about how we might address the possibility that our technology might grow beyond our capacity to control it. Asimov felt that this catastrophe could possibly be averted if we designed a moral code into our creations. He intended that the basic premise under these laws (to bring no harm to humans) would be enough to ensure that an artificial intelligence could be acceptable to the general public [Asimov 1942]. Codifying morality into our creations would provide a counteraction to evolution's basic tendency to push a goal driven AI system to the point of being dangerous to humanity [Yudkowsky, 2006]. The obvious hole in this approach is that it relies on our own self-discipline as a society to implement. Historically, in the context of weaponization, war, and human conquest, we as a species are incapable of imposing remotely similar limitations on ourselves.

The Concept of a Friendly AI

Beyond simply coding the rules of "do no harm" into the machines, another group, "The Singularity Institute" proposes that our efforts should be focused on building the capability to embed human sensibilities into an AI creation. Termed "Friendly AI", this concept addresses the much more immediate (and likely) problem of when one of our creations gets out of control while performing a simple or benign task. Proponents of Friendly AI are less concerned with humanity building something that could deliberately cause harm, and more concerned with the possibility that we will develop something that is completely indifferent to our existence [Yudkowsky, 2006]. Eliezer Yudkowsky makes a case for this with his description of how a system with the ability to use molecular nanotechnology to improve itself, is given the task of solving the Riemann hypothesis. In his description, the system eventually escapes our ability to control it and begins converting all matter in the solar system in its drive to acquire the resources necessary to solve the problem. The system only cares about achieving its goal. Unfortunately, the pursuit of this goal requires that it acquire an increasing amount of limited resources. Acquisition of these resources is in direct conflict with our own existence, and ultimately the system destroys itself and the very beings that asked it the original question [Yudkowsky, 2006].

Friendly AI arguments assert that in order to protect humanity from the above problem, researchers need to ensure that every system they build consists of goals that are "human friendly". This stems from the assumption that a Superintelligence, given enough time, will be able to achieve whatever goal it seeks. Therefore the approach is to ensure that these goals remain fundamentally benevolent within the context of human society. The philosophy has determined that the following requirements are necessary for Friendly AI to be effective [Yudkowsky, 2006]:

1. *Friendliness* - that an AI feel sympathetic towards humanity and all life, and seek for their best interests.

2. *Conservation of Friendliness* - that an AI must desire to pass on its value system to all of its offspring and inculcate its values into others of its kind.
3. *Intelligence* - that an AI be smart enough to see how it might engage in altruistic behavior to the greatest degree of equality, so that it is not kind to some but more cruel to others as a consequence, and to balance interests effectively.
4. *Self-improvement* - that an AI feel a sense of longing and striving for improvement both of itself and of all life as part of the consideration of wealth, while respecting and sympathizing with the informed choices of lesser intellects not to improve themselves.
5. *First mover advantage* - the first goal-driven general self-improving AI "wins" in the memetic sense, because it is powerful enough to prevent any other AI emerging, with might compete with its own goals.

In addition, it is also the belief of this group that the task of developing the goals at the heart of "Friendly AI" cannot be developed by a single or small group of humans. Instead, Yudkowsky asserts via his "Coherent Extrapolated Volition" model that the common will of humanity converges on coherent set of goals. Determining this set of goals will be the work of an early Seed AI project programmed to observe and analyze human nature [Yudkowsky, 2004]

Embedding Morality: Social Contracts

Further extending this concept is the idea of designing a social contract into the machine. Bill Hibbard in his essay "The Technology of Mind and a New Social Contract" implies that those with the means to build a Seed AI system will be wealthy corporations or governments. Neither of these entities have any incentive to incur the additional expense of building "Friendliness" into their systems. It is quite conceivable that a Superintelligence could be built to serve the purposes of its creator only, and not be embedded with any philanthropic values [Bostrom, 2003]. Hibbard goes on further to note that the concepts of both Friendly AI and Asimov's Laws of Robotics are valid, but cannot be achieved by means of a purely technical solution because they are inherently ambiguous [Hibbard, 2008]. Amplifying this concept is Kurzweil, who stresses the point that a "greater intelligence will always find a way to circumvent measures that are the product of a lesser intelligence" [Kurzweil, 2005].

Hibbard's approach instead is to leverage the fundamental capabilities of a "mind" from the perspective of how it is able to improve itself through the concept of "reinforcement learning" through its interaction with a greater society. He defines this as a "brain" encoded with a basic set of values that attempts to maximize these values through simple trial and error. Reinforcement learners therefore have the following properties [Hibbard, 2008]

1. Reinforcement values to be increased or decreased - these are the basic motives of behavior
2. Algorithms for learning behaviors based on reinforcement values.
3. A simulation model of the world, itself learned from interactions with the world (the reinforcement value for learning the simulation model is accuracy of prediction).
4. A discount rate for balancing future versus current rewards (people who focus on current rewards and ignore the future are generally judged as not very intelligent).

The properties above help quantify the options that are important to the development of a "mind", and help us decide how to determine the values and establish the discount rate for the future rewards [Hibbard, 2008]. An evolutionary trade-off exists between the self and the society. Choices made by an individual that favor that individual at the expense of the society tend to achieve short term results only as they are typically frowned on by the greater group. Conversely, a purely selfless choice will be rewarded by society for the long term, but runs the risk of eliminating the entity and therefore its ability to benefit from the reward. Humans have evolved a series of compromises to this trade-off. The positive (pleasure) and negative (pain) reinforcement behind our own emotions has evolved to help us introspectively evaluate the choices we make.

Minsky extends this by adding to it the concept of the “imprimer”. In his discussions he asserts that there is additional weight put on the mind’s reinforcement when specific members of society perform it. An example of this is the dissatisfaction a parent has for a child’s behavior tends to have more influence in modifying that child’s behavior than the same level of dissatisfaction shown by a complete stranger [Minsky, 2006]. It would be wise to consider this in our model for reinforcement learning as well.

At this point, a social contract is necessary to discipline our efforts toward the creation of motives and the establishment of values within our intelligent machines. These motives must weigh the greater good of society against the immediate selfish goals of the individual. Reinforcement learning then becomes the mechanism behind strengthening these values, thereby influencing the future behavior of the machine. Because all of human society stands to benefit (or suffer) as a result of the super intelligence, it demands that the

contract be agreed upon and enforced at the political levels across the world. AI research by governments involved in weapons research should be obligated and held accountable for adhering to the concepts in the social contract. This is very similar to how chemical, nuclear, and biological weapons are treated today.

IV. ROADMAP FOR THE FUTURE

Perhaps the best first start is to begin raising the public's awareness of the issue. Socializing the concept that there is a very real possibility of an approaching singularity may be enough to get humanity talking about how to prepare for it. Once our collective minds begin working on the problem, a solution for coexistence may emerge. At a minimum, as a species, we will be able to come up with a plan so that we are not caught completely off guard.

Clear Justification to Proceed

Like the controversial chemical, nuclear, and biological warfare technologies of the previous century, there is a strong justification to continue the development of a Superintelligence. Humanity has always had a driving need to acquire knowledge and understanding. The application of this has given us fantastic technological advances. A Superintelligence has the potential to be an indispensable tool in every field of science. The ability to wield incredibly complex amounts of information could lead to advances such as immortality, and interstellar space flight [Bostrom, 2003]. While the good reasons alone might well be enough justification, the bad ones make it impossible to ignore. Humanity is also capable of great selfishness and destruction. A technology as powerful as a super-intelligence in the wrong hands could easily lead to the end of humanity. It is therefore the duty of open societies to continue focusing research efforts in the direction of a Superintelligence to ensure they are able to get there first.

Transparency and Regulation

In an ideal world, transparency would be the norm and all ideas and concepts would be freely accessible to anyone. Unfortunately, we are compelled to develop new technologies in secret. It is reasonable to assume that only a government or wealthy corporation would be able to summon the massive resources required to develop a Superintelligence. Past history makes it obvious to also assume there will be very little transparency during the early stages of development. It is therefore necessary that our society ensure there are adequate safeguards in place to protect the greater public. Safeguards to regulate containment of the technology, a definition of the concepts and desires behind the social contract, and the appropriate regulatory statutes to ensure these wishes are followed appropriately. Again, the framework and a potential starting point for the construction of these safeguards might be found by looking to the past models of chemical, nuclear, and biological technology management.

Toward a Formal Definition of Machine Rights

The current ethical consensus is that legal rights are not assignable to non-human animals because it is not possible to enter into a moral contract with them. Given the current state of the art, machines would therefore be a logical extension of this concept because no machine currently exists that possesses the ability to actually understand the contents of a contract. However, the caveat with this is within the word "currently". The looming realization of the Technological Singularity would obviously negate this reasoning and force our society to rethink this stance.

Perhaps the best approach is to clarify and further define the concept of assigning legal rights as a condition of an entity's ability to achieve an understanding of those rights. Put another way; does the machine "care" if it has rights or not? If we couch it this way, we can cover the possibility that the emergence of a self-aware super-intelligence will not catch our society off guard. Instead, we would be prepared to welcome such a machine based solely on the facts that it "wants to", and that it actually "understands what that means." By defining legal rights in these terms, we open up the possibility of automatically granting machine rights if the necessity ever arises.

Unfortunately, significant obstacles exist in realizing this concept. As Benjamin Soskis points out, legal rights to date have never been granted in abstraction. To do so would be problematic because of the potential ambiguity it may cause. In addition to this, we have our own human prejudices to overcome. Many in this subject area insist that no matter how advanced a machine is, it will never have an “intrinsic moral worth” [Soskis, 2005]. The rationale behind this is deeply rooted in religious superstition about the origin of the “soul.” Despite being unfounded in completely rational thought, it is still reason enough for many people to outright dismiss the potential a machine will ever qualify for any kind of natural rights. The reconciliation between this way of thinking and the reality of the approaching Technological Singularity must be achieved before society can begin acceptance of machine intelligence.

Scientists and researchers must be cognizant of the risk being created by continuing to push the boundaries of modern research. Unfortunately, it is not realistic to think we will ever be able to universally impose a

system of ethics on research and design (e.g. there will always be weapons designers). While this is a noble goal that is worthy of attempt, it should not be our only strategy. An additional focus should also be implemented toward making the rules of membership into human society compatible with our creations. Historically, it has always been easier to assimilate new cultural changes after we have made steps to prepare ourselves for the transition. Kurzweil feels that as we develop an increasingly symbiotic relationship with our machines, this will be a natural transition [Kurzweil 1999]. This is an optimistic view that may breed some unwanted complacency. Catching humanity unaware and unprepared would be the worst scenario, as it is conceivable we would immediately ostracize this new class of citizen. The resentment this would foster, coupled with the expectation that this entity could out-think us would immediately set our species down the path to evolutionary extinction.

V. CONCLUSION

Bill Joy warns us in no uncertain terms that the future doesn't need us. The evolution of information is reaching a point where biological minds may not be able to keep up with artificial ones. All evidence across history points to the eventual demise of the inferior species at the hands of the superior. The warning is simply that if machines are permitted to make their own decisions, we will lose the ability to influence those decisions and eventually be completely at their mercy [Joy, 2003]. This dependency can arise simply from society yielding the operation of complex systems to machine intelligence until eventually no one but the machines understand how the systems function. The ultimate implication of this is that humanity no longer serves a useful purpose, and is therefore no longer needed.

Given the accelerating pace and direction of technological advancement, it is reasonable to assume that in the near future, our species will create the ability to automate the manipulation of vast amounts of information. Couple this with the continuing advancements in other fields such as self-replication and nanotechnology, and the environment seems to be developing for the creation of a Superintelligence. Slowing the pace of technology is not an option, and neither is ignoring the issue. We as a society must begin thinking about how we will face this challenge and put the plans in place that ensure we remain relevant in the future we are creating.

REFERENCES

[1] Asimov, I., (1942) "Runaround". *Astounding Science Fiction*, March 1942.

[2] Bostrom, N., (2003) "Ethical Issues in Advanced Artificial Intelligence," *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, Int. Institute of Advanced Studies in Systems Research and Cybernetics, vol. 2, ed. I., pp. 12–17, 2003.

[3] Bostrom, N., (1998) "How Long before Superintelligence?", *International Journal of Future Studies*, vol. 2, 1998.

[4] Čapek, K., (1920) "Rossum's Universal Robots", *Toward the Radical Center*, North Haven, CT: Catbird Press, 1990.

[5] Hibbard, B., (2008) "The Technology of Mind and a New Social Contract," *Journal of Evolution and Technology*, Institute for Ethics and Emerging Technologies, Vol. 17, Issue 1, pp 13-22, January 2008.

[6] Joy, B., (2003) "Why the Future Doesn't Need Us," *Wired Magazine*, Wired Digital, Inc., 2003, retrieved August 6, 2011, from http://www.wired.com/wired/archive/8.04/joy_pr.html

[7] Kurzweil, R., (1999) *The Age of Spiritual Machines*. New York, NY: Penguin Books, 1999.

[8] Kurzweil, R., (2001) "The Law of Accelerating Returns", *KurzweilAI Essays*, retrieved August 6, 2011, from http://www.kurzweilai.net/the-law-of-accelerating-returns.

[9] Kurzweil, R., (2005) *The Singularity is Near: When Humans Transcent Biology*. New York, NY: Penguin Books, 2005.

[10] Levene Gera, D., (2003) *Ancient Greek Ideas on Speech, Language, and Civilization*, New York, NY: Oxford University Press, November 2003.

[11] Minsky, M., (2006) *The Emotion Machine*, New York, NY: Simon & Schuster, 2006.

[12] Russell, S. J., & Norvig, P., (2010) *Artificial Intelligence: A Modern Approach* (3rd ed.). Upper Saddle River, New Jersey: Pearson Education, Inc., 2010.

[13] Soskis, B., (2005) "Man and the Machines: It's Time to Start Thinking About How We Might Grant Legal Rights to Computers," Legal Affairs, Jan-Feb 2005, retrieved August 6, 2011 from http://www.legalaffairs.org/issues/January-February-2005/feature_sokis_janfeb05.msp.

[14] Taddei, M., (2007) *Leonardo da Vinci's Robots*, London, United Kingdom: Citigate Publishing Ltd., 2007.

[15] Yudkowsky, E., (2006) "Artificial Intelligence as a Positive and Negative Factor in Global Risk, " unpublished., *Global Catastrophic Risks*, Singularity Institute for Artificial Intelligence August 31, 2006.

[16] Yudkowsky, E., (2004) "Coherent Extrapolated Volition", *Singularity Institute for Artificial Intelligence*, retrieved August 6, 2011 from http://singinst.org/upload/CEV.html.

[17] Yudkowsky, E., et. al., (2001) "General Intelligence and Seed Artificial Intelligence: Creating Complete Minds Capable of Open-Ended Self-Improvement," *Singularity Institute for Artificial Intelligence*, retrieved August 6, 2011 from http://singinst.org/ourresearch/publications/GISAI/index.html