

THE SINGULARITY INSTITUTE FOR ARTIFICIAL INTELLIGENCE

2007 Overview & 2006 in Review

SIAI Mission & Objectives

In the coming decades, humanity will likely create a powerful AI. The Singularity Institute for Artificial Intelligence exists to confront this urgent challenge, both the opportunity and the risk.

SIAI is a not-for-profit research institute based in Palo Alto, California, with three major goals: furthering the nascent science of safe, beneficial advanced AI (self-improving systems) through research and development, research fellowships, research grants, and science education; furthering the understanding of its implications to society through the AI Impact Initiative and annual Singularity Summit; and furthering education among students to foster scientific research.

To learn more about SIAI, see online “What is the Singularity?”, “Why Work Toward the Singularity?”, and “Artificial Intelligence as a Positive and Negative Factor in Global Risk”.

Progress to Date in 2007

National Medal of Technology winner Ray Kurzweil joined the SIAI Board of Directors. Dr. Ben Goertzel became SIAI Director of Research to expand the Institute’s research program. Bruce Klein became SIAI Director of Outreach to grow our education and outreach efforts, beginning with the Singularity Summit 2007. The SIAI R&D Program was created and pilot funding began being sought for Phase I. The Singularity Summit 2007: Artificial Intelligence and the Future of Humanity was announced, scheduled for September 8-9 at the Palace of Fine Arts Theatre in San Francisco, CA. The Singularity Summit at Stanford DVD and online media were released. The \$400,000 Matching Challenge and SIAI Donor Network were announced to build the Singularity Institute’s donor base for support of research, education, and outreach expansion in 2007.

Looking Back: 2006 in Summary

SIAI co-sponsored the Singularity Summit at Stanford, conceived and organized by Executive Director Tyler Emerson, in collaboration with Professor Todd Davies of the Symbolic Systems Program at Stanford, Ray Kurzweil, Yonah Berwaldt, and Michael Jin. SIAI created the structure for the SIAI Research Grant Program (seed funding is presently being sought), a new program in partnership with the Future of Humanity Institute at Oxford. SIAI Research Fellow Eliezer

Yudkowsky released “AI as a Positive and Negative Factor in Global Risk” and “Cognitive Biases Potentially Affecting Judgment of Global Risks,” publications forthcoming from Oxford University Press in the edited volume *Global Catastrophic Risks*. Yudkowsky presented at the Bay Area Future Salon, Singularity Summit, and AGI Workshop. SIAI raised \$200,000 through a Challenge Grant backed by Peter Thiel, following which Marcello Herreshoff began as a Research Associate, and Allison Taguchi joined as Director of Development. Neil Jacobstein, Dr. Stephen Omohundro, Dr. Barney Pell, and Peter Thiel joined the SIAI Board of Advisors.

SIAI Team



SIAI Executive Director Tyler Emerson is responsible for overseeing and scaling the Singularity Institute's operations. His focus includes administration and management, organizational planning, fundraising, writing, speaking, event organizing, and talent outreach and recruitment. He is Chair of the Singularity Summit II: Artificial Intelligence and the Future of Humanity. He is organizing and overseeing the AI Impact Initiative, in collaboration with Dr. Stephen Omohundro, to bring together a multidisciplinary body of experts to examine the critical questions of advanced AI's potential impact on humanity. He recently created the SIAI Research Grant Program, in partnership with the Future of Humanity Institute at Oxford, to foster scientific research on the Foundations of Reflection, Friendly AI, Existential Risks, and Human Enhancement. Previously, he was the main organizer of the Singularity Summit at Stanford University, the first academic symposium on the singularity scenario, which brought together 1300 people and 10 leading voices to explore the future of human and machine cognition. Previously, he worked with the Acceleration Studies Foundation, co-organizing the first Accelerating Change Conference in 2003; and the Foresight Nanotech Institute, the premier nanotechnology public interest nonprofit.



SIAI Director Ray Kurzweil, CEO of Kurzweil Technologies, has been described as "the restless genius" by the Wall Street Journal, and "the ultimate thinking machine" by Forbes. Inc. Magazine ranked him #8 among entrepreneurs in the US, calling him the "rightful heir to Thomas Edison," and PBS included him as one of the 16 "revolutionaries who made America," along with other inventors of the past two centuries. As one of the leading inventors of our time, Ray has worked in such areas as music synthesis, speech and character recognition, reading technology, virtual reality, and cybernetic art. He was the principal developer of the first omni-font optical character recognition, the first print-to-speech reading machine for the blind, the first CCD flat-bed scanner, the first text-to-speech synthesizer, the first music synthesizer capable of recreating the grand piano and other orchestral instruments, and the first commercially marketed large-vocabulary speech

recognition. All of these pioneering technologies continue today as market leaders. His website, KurzweilAI.net, has over one million readers. Among his many honors, he is the recipient of the \$500,000 MIT-Lemelson Prize, the world's largest for innovation. In 1999, he received the National Medal of Technology, the nation's highest honor in technology, from President Clinton. In 2002, he was inducted into the National Inventor's Hall of Fame, established by the US Patent Office. Ray has also received twelve honorary Doctorates and honors from three U.S. presidents. His books include *The Age of Intelligent Machines*, *The Age of Spiritual Machines*, and *Fantastic Voyage: Live Long Enough to Live Forever*. Three of his books have been national best sellers. His latest, published by Viking, is *The Singularity is Near: When Humans Transcend Biology*.



SIAI Director of Research Dr. Ben Goertzel oversees the direction of the Institute's research program. He has over 70 publications, concentrating on cognitive science and AI, including *Chaotic Logic*, *Creating Internet Intelligence*, *Artificial General Intelligence* (edited with Cassio Pennachin), and *Hidden Pattern*. He is chief science officer and acting CEO of Novamente, a software company aimed at creating applications in the area of natural language question-answering. He also oversees Biomind, an AI and bioinformatics firm that licenses software for bioinformatics data analysis to the NIH's National Institute for Allergies and Infectious Diseases and CDC. Previously, he was founder and CTO of Webmind, a 120+ employee thinking-machine company. He has a Ph.D. in mathematics from Temple University, and has held several university positions in mathematics, computer science, and psychology, in the US, New Zealand, and Australia.



SIAI Research Fellow Eliezer Yudkowsky is one of the world's foremost researchers on Friendly AI and recursive self-improvement. He created the Friendly AI approach to AGI, which emphasizes the importance of the structure of an ethical optimization process and its supergoal, in contrast to the common trend of seeking the right fixed enumeration of ethical rules a moral agent should follow. In 2001, he published the first technical analysis of motivationally stable goal systems, with his book-length *Creating Friendly AI: The Analysis and Design of Benevolent Goal Architectures*. In 2002, he wrote "Levels of Organization in General Intelligence," a paper on the evolutionary psychology of human general intelligence, published in the edited volume *Artificial General Intelligence* (Springer, 2006). He has two papers forthcoming in the edited volume *Global Catastrophic Risks* (Oxford, 2007), "Cognitive Biases Potentially Affecting Judgment of Global Risks" and "Artificial Intelligence as a Positive and Negative Factor in Global Risk."



SIAI Director of Corporate & Foundation Relations Allison Taguchi oversees all corporate and foundation giving for the Singularity Institute. She comes to SIAI with over 12 years of fund development experience at research institutes, universities, government agencies and nonprofit organizations. Some of the businesses she worked with in the past include: Rushford Nanotech Laboratory, Dept. of Defense, Oakland Military Institute, and University of Hawaii Biotech Research Center.



SIAI Director of Outreach Bruce Klein is focused on raising the awareness and understanding of the Singularity Institute's mission to foster beneficial artificial general intelligence. He is an entrepreneur and social activist, presently serving as president of Novamente, and previously known for creating the Immortality Institute. In 2004, he edited *The Scientific Conquest of Death*, a volume devoted to examining the biological and philosophical arguments against dramatically extending the healthy human lifespan, with 18 contributors, including Michael Rose, Aubrey de Grey, and Michael West. In the same year, the tragic death of Klein's mother led him to create a documentary film to promote the Immortality Institute's mission to overcome the "blight of involuntary death." He traveled the US, interviewing scientists and futurists on the subject of life extension, leading to the film "Exploring Life Extension." He unveiled the film at the first Life Extension Conference, organized in 2005 by Klein and his wife Susan-Fonseca Klein.

SIAI Research Associate Marcello Herreshoff works directly with Research Fellow Eliezer Yudkowsky on Friendly AI theory and related mathematics. Herreshoff deferred his acceptance at Stanford University for a year to work at SIAI. He has an extensive background in mathematics and computer science, having amassed numerous awards and prizes during both local and national competitions. In computer science, he participated in the United States of America Computing Olympiad (USACO) and became one of the sixteen national finalists to participate in a rigorous training camp to prepare the finalists to compete in the International Computer Olympiad. Aside from his achievements in computer science, Herreshoff recently presented at the Twelfth Conference on Fibonacci Numbers and their Applications. He had discovered a novel proof using combinatorics to prove a previously known identity involving Fibonacci Numbers.

SIAI Advisors

Dr. Nick Bostrom, Founder and Director, Future of Humanity Institute at Oxford

Dr. Aubrey de Grey, Chairman and Chief Science Officer, Methuselah Foundation

Neil Jacobstein, Chairman and CEO, Teknowledge Inc.

Dr. Stephen Omohundro, Founder and President, Self-Aware Systems

Dr. Barney Pell, Founder and CEO, Powerset Inc.

Christine Peterson, Co-Founder and Vice President, Foresight Nanotech Institute

Peter Thiel, Founder and Managing Member, Clarium Capital Management

New: The Singularity Summit 2007

Theme: Artificial Intelligence and the Future of Humanity

When: Friday and Saturday | September 8 - 9 | 9:00 AM - 6:00 PM

Where: Palace of Fine Arts Theatre | San Francisco, CA

Audience: Scientists, technologists, c-level execs, foundation heads, researchers, entrepreneurs, students, forecasters, programmers, philanthropists, cultural leaders, VCs, bloggers, geeks, press

Speakers: Rodney Brooks (MIT AI Lab Director), Peter Norvig (Google Director of Research), Paul Saffo (Institute for the Future Roy Amara Fellow), Barney Pell (Powerset CEO), and more

In recent years, several scientists and authors have argued there is a significant chance advanced AI will be developed in a few decades. Similar claims, however, have been made over the past 50 years. What is different now? The Singularity Summit 2007 will bring together 20 premier thinkers to examine whether we are really nearing a turning point toward powerful intelligence.

Last year, the Singularity Summit at Stanford University, the first academic symposium explicitly focused on the singularity scenario, brought together 1300 people and 10 speakers to explore the future of human and machine cognition, including Ray Kurzweil, Dr. Douglas R. Hofstadter, and Dr. Sebastian Thrun. Press coverage included the San Francisco Chronicle, ZDNET, and Business 2.0. Presentations can be viewed online or downloaded at <http://new.singinst.org/media>.

The Singularity Summit 2007 at the Palace of Fine Arts Theatre in San Francisco will be an important exploration of what may become a historical moment in time – a window of opportunity to affect how the world moves forward with a powerful new technology.

New: SIAI \$400,000 Matching Challenge

Immediate Deadline: SIAI must raise \$200,000 by the end of Monday, May 21st, 2007, to receive an additional \$50,000 commitment from Peter Thiel, PayPal Founder and SIAI Advisor.

Donations can be made to the Matching Challenge at <http://www.singinst.org/challenge/>.

SIAI has built the infrastructure to expand our research, education, and outreach in 2007. Thanks to the generous support of matching fund donors, including Peter Thiel, we have announced the \$400,000 Matching Challenge, which, if successful, will help underwrite our expansion this year.

Every dollar you donate will be matched dollar-for-dollar by our matching donors until **July 6th**.



I have recently joined the SIAI Board of Directors, because the Singularity Institute is playing a critical role in advancing humanity's understanding of the profound promise and peril of strong AI. I hope you will consider supporting the Institute. Your contribution will help support the new R&D program, expand fundraising and educational outreach, and ensure a successful Singularity Summit II. – Ray Kurzweil, Founder and CEO, Kurzweil Technologies, SIAI Director

What the 2007 Challenge Will Support

- R&D Program (Postdoctoral Fellow, Ph.D. Scholar, Research Programmer)
- The Singularity Summit II (Sep 8-9, Palace of Fine Arts Theatre, San Francisco, CA)
- Research Fellowship Program & Fund (to Support Present & Future Research Fellows)
- Fund Development (Roundtable Dinners, Special Events, Foundation Grant Proposals)
- Educational Outreach & Marketing (Site, Blog, Newsletter, Media, Online Marketing)

The Singularity Institute for Artificial Intelligence is a 501(c)(3) nonprofit organization. Your donation will be tax deductible to the fullest extent possible in the United States and Canada.



The ideas of a singularity scenario and Friendly AI are powerful, but more importantly, relevant now. I support SIAI because they are making unique contributions to these critical areas of knowledge. I have pledged \$200,000 in matching funds to the matching challenge to help grow the Singularity Institute, with an additional \$50,000 available if half of the challenge is matched in the first 45 days. I hope you will join me as a donor, and support this important organization. – Peter Thiel, Clarium Capital Management Founder, PayPal Founder & Former CEO, SIAI Advisor

New: SIAI R&D Program

Our mission is to handle the challenge of powerful AI, both the opportunity and risk. One of our paths toward this is research and software development. We have three aims with our research:

- Achieve a fundamental understanding of the problems underlying the creation of safe, beneficial AI with powerful general intelligence.
- Provide the AI community with conceptual, mathematical, and software tools that they can use to move and accelerate their R&D in the direction of safe, beneficial AI.
- Work directly toward the development of safe, beneficial, general AI.

Our goals are different in emphasis and focus than contemporary academic/industry AI research communities. The SIAI R&D Program differs from a typical AI R&D program in two key ways:

- Our focus on general intelligence, as opposed to narrow AI software, which is designed to solve specialized problems, such as chess-playing, flight navigation, or fraud detection.
- Our focus on safety and beneficial use. One of our basic positions is that these qualities must be placed at the foundation of work on general AI (Artificial General Intelligence, or AGI), rather than tacked on at the end of the theory, design, and development process.

Phases

The SIAI R&D Program consists of two phases:

Phase I, 2007 – 2012 (approx.):

Develop a unified mathematical and conceptual framework for studying safe, beneficial AGI; and create software tools and technologies to assist with AGI design, engineering, and instruction.

Phase I mathematical and conceptual research will include:

- The use of probability, decision, and algorithmic information theory to study AGI ethics.
- The development of a shared ontology for discussing cognitive concepts important for describing AGI designs and their behaviors.

Phase I software tools and technologies will include:

- A simulation world for AGIs.
- A collection of IQ/ethical test scenarios for AGIs, implemented in the simulation world.
- A language for communicating with AGIs about ethics, etc. with minimal ambiguity.
- Knowledge resources allowing AGIs to more easily represent and reason about their own structure and dynamics.
- Key cognitive technologies designed specifically to serve as flexible components within safe, beneficial AGI architectures.

Phase II, 2012 – : If viable, design and engineer safe, beneficial AGI, utilizing the theories, tools, and technologies developed in Phase I.

Both of our phases will likely involve a combination of in-house research and intensive collaboration with researchers in academia and industry.

SIAI R&D Program Research Areas

Our Phase I research divides into three categories:

- Theoretical Research
- Tools and Technologies
- System Design and Implementation (primarily Phase II)

Theoretical Research

Research Area 1: Mathematical Theory of General Intelligence

Our research in this area will focus on using algorithmic information theory and probability theory to formalize the notion of general intelligence, specifically ethical general intelligence. Important work in this area has been done by Marcus Hutter, Jürgen Schmidhuber, Shane Legg, and others, as well as by our team; but this work has not yet been connected with pragmatic AGI designs. Meeting this challenge is one of our major goals going forward. Specific focus areas within this domain include:

- **Mathematical Formalization of the "Friendly AI" Concept.** Proving theorems about the ethics of AI systems, an important research goal, is predicated on the possession of an appropriate formalization of the notion of ethical behavior on the part of an AI. And, this formalization is a difficult research question unto itself.
- **Implications of Algorithmic Information Theory for the Predictability of Arbitrarily Intelligent AIs.** In 2006, Shane Legg made an interesting, ultimately failed attempt to prove algorithmic information theoretic limitations on the possibility of guaranteeing ethical behavior on the part of future AIs. This line of research however has significant potential for future exploration.
- **Formalizing the Concept of General Intelligence.** Shane Legg and Marcus Hutter published a paper in 2006 presenting a formal definition of general intelligence. Their work is excellent but can be extended in various ways; in particular, work is needed on connecting these ideas with practical intelligence tests for AGIs.

- **Reflective Decision Theory: Extending Statistical Decision Theory to Strongly Self-Modifying Systems.** Statistical decision theory, as it stands, tells us little about software systems that regularly make decisions to modify their own source code in radical ways. This deficit must be remedied if we wish to formally understand self-modifying AGI systems, their potential dangers, and potential routes to ensuring their long-term safety and beneficialness.
- **Dynamics of Goal Structures Under Self-Modification.** Under what conditions will an AGI system's internal goal structure remain invariant as the system self-modifies? Supposing that one of the system's top-level goals is precisely this sort of goal-system invariance – nevertheless, that is clearly not enough to guarantee invariance. Additional conditions are needed, but the nature of these conditions has not been seriously investigated. This is a deep mathematical issue in the dynamics of computational intelligence, with obvious critical implications for the creation of stably beneficial AGI.

Research Area 2: Interdisciplinary Theory of AGI

One of our objectives in this area is to create a systematic framework for the description and comparison of AGI designs, concepts, and theories. We will also make selective contributions relevant to the practicalities of creating, engineering, and understanding real-world AGI systems.

- **Mind Ontology: A Standardized Language for Describing AGI Systems and Related Concepts.** One of the issues holding back AGI progress is that different researchers often use different languages to discuss the same things. One solution is to agree upon a standard ontology of AGI-related concepts. An initial draft of such an ontology exists, but needs extension and refinement. A description of current AGI designs and cognitive neuroscience knowledge in terms of the common ontology also needs to be undertaken.
- **AGI Developmental Psychology.** Once an AGI is engineered, it will have to be taught: the first AGI will likely be more like an artificial baby than an artificial adult. Current theories of developmental psychology are focused on human psychological development. However, if AGI development begins by creating "baby AGI's" and teaching them gradually, we will need a theory of AGI developmental psychology to guide our work. Recent theoretical work by Ben Goertzel and Stephan Vladimir Bugaj took a small step in this direction by connecting Piagetan developmental psychology with the theory of uncertain inference; but considerably more research is required. One of the key issues here is the interdependence of ethical development with cognitive development, which is only moderately understood in humans, and will likely be quite different in AGIs.

Research Area 3: AGI Ethical Issues

A central view of our research team is that ethical issues must be placed at the center of AGI research, rather than tacked on peripherally to AGI designs created without attention to ethical considerations. Several of our focus areas have direct implications for AGI ethics (particularly the investigation of goal system stability), but we also intend to heavily investigate several other issues related to AGI and ethics, including:

- **Formalizing the Theory of Coherent Extrapolated Volition.** Research Fellow Eliezer Yudkowsky has proposed "coherent extrapolated volition" (CEV) as a way of arriving at a top-level supergoal for an AI that represents the collective desires of a population of individuals. While fascinating, the idea has only been presented informally, and a mathematical formalization seems necessary so that its viability can be assessed. For example, it is of interest to try to articulate formally the conditions under which the CEV of a population of individual agents, appropriately defined, will exist. This may depend on the coherence versus divergence of the beliefs or mind-states of the individuals.
- **Framework for Formalizing Desired Beneficial Outcomes.** To create safe and beneficial AI systems, we must have a clear vision of what constitutes a beneficial outcome. The recently developed science of Positive Psychology is making great strides in understanding elements that promote human happiness. Political philosophy has studied a wide variety of approaches to structure "the good society" in a way that maximizes the benefits to its citizens. We will work toward creating a framework which formalizes these kinds of insights so that they can be considered for AI goal systems.
- **Decision-Theoretic and Game-Theoretic Foundations for the Ethical Behavior of Advanced AIs.** Microeconomics and decision theory study the nature of individual preferences and their influence on behavioral outcomes. Game theory is the core mathematical theory of decision making by interacting agents. We will use these tools to analyze the likely behavior of alternative models for the safe deployment of advanced self-modifying AIs. The preferences of an agent together with the behavior of other agents in its environment determine the actions it will take. We must design the preferences of agents so that their collective behavior produces the results we desire and is stable against internal corruption or external incursion.

Tools & Technologies

This is a broad but critical area. One thing that has delayed AGI research is the scarcity of useful software tools, including for measuring ethicalness. In order to serve our R&D and the R&D of external researchers, the creation of a suite of relevant software tools will be invaluable.

Research Area 4: Customization of Existing Open-Source Projects

Our initial work in this area will focus on customizing and further developing existing open-source software projects. There are valuable, preexisting projects moving slowly due to lack of funding, which can be morphed into specific tools for aiding the creation of safe, beneficial AGI. Three main examples are the AGISim simulation world project, the Lojban language for human-machine communication, and the Mizar mathematics database.

Like any complex engineering challenge, building an AGI involves a large number of tools, some of which are quite complex and specialized. One delay of progress in AGI is the lack of appropriate tools. Each team must develop their own, which is time-consuming and distracts attention from the actual creation of AGI designs and systems. One of the key roles SIAI can play going forward is the creation of robust tools for AGI development, to be utilized in-house and by the AGI research community at large.

- **AGISim, a 3D Simulation World for Interacting with AGI Systems.** AGISim is an open-source project in alpha release. It is usable, but still needs more coding work done. A related task, of significant use to robotics researchers, is the precise simulation of existing physical robots within AGISim. AGISim also plays a key role in some of the AGI IQ/ethics evaluation tasks to be described below.
- **Lojban: A Language for Communicating with Early-Stage AGIs.** Lojban is a constructed language with hundreds of speakers, based on predicate logic. Thus, it is particularly suitable for communication between humans and AGIs. A Lojban parser exists, but needs to be modified to make it output logic expressions, which will then allow Lojban to be used to converse with logic-based AGI systems. This will allow communication with a variety of AI systems in a human-usable yet relatively unambiguous way, which will be valuable for instructing AGI systems, including ethical behavior instruction.
- **Translating Mizar to KIF.** Mizar is a repository of mathematical knowledge, available online but in a complex format that is difficult to feed into AI theorem-proving systems. In six months, a qualified individual could translate Mizar to KIF, a standard predicate logic format, which would enable its use within theorem-proving AI systems, a crucial step toward AGI systems that can understand themselves and the algorithms utilized within their sourcecode.

Research Area 5: Design and Creation of Safe Software Infrastructure

Some key areas of tool development are not adequately addressed by any current open-source project, for example, the creation of programming languages and operating systems possessing safety as built-in properties. SIAI researchers would not be able to complete such large, complex

projects on their own, but SIAI can potentially play a leadership role by articulating detailed designs, solving key conceptual problems, and recruiting external partners to assist with engineering and testing.

- **Programming Languages that Combine Efficiency with Provability of Program Correctness.** In the interest of AGI safety, it would be desirable if our AGI software programs could be proved to correctly implement the software designs they represent. However, currently, there is no programming language that both supports proof-based program correctness checking, and is sufficiently efficient in terms of execution to be usable for pragmatic AGI purposes. Creating such a programming language framework will require significant advances in programming language theory.
- **Safe Computer Operating Systems.** Is it feasible to design a provably correct operating system? In principle, yes, but this task would likely require a programming language that combines efficiency with provable correctness, as well as several interconnected breakthroughs in operating systems theory. Creating a version of Unix in a programming language that supports provable correctness would be a start, but there are many issues to be addressed. This is a research topic that requires close collaboration between a mathematician and an experienced operating systems programmer.

Research Area 6: AGI Evaluation Mechanisms

The creation of safe, beneficial AGI would be hastened if there were well-defined, widely-accepted means of assessing general intelligence, safety, and beneficialness. The provision of such means of assessment is a tractable task that fits squarely within the core mission of SIAI.

A few comments regarding AGI intelligence testing must be inserted here, as general context. IQ tests are a controversial but somewhat effective mechanism for assessing human intelligence. Narrow AI software is evaluated by a variety of mechanisms appropriate to the various domains in which it operates. AGI software, on the other hand, is not currently associated with any generally accepted evaluation mechanism. The Turing Test and variations purport to assess the effectiveness of AGI systems at emulating human intelligence, but have numerous shortcomings: not all AGI systems will necessarily aim at the emulation of human intelligence; and furthermore, these tests do not provide any effective way of assessing the continual progress of AGIs toward more and more general intelligence. The Loebner Prize, a chat-bot contest, purports to assess the progress of AI systems toward general intelligence in a conversational fluency context, but its shortcomings have been well documented. It is with this background in mind that we propose to devote some effort to the creation of intelligence evaluation mechanisms focused specifically on AGI. We do not expect this to lead to any single, definitive "AGI IQ test," but rather to a host of evaluation mechanisms that are useful to AGI researchers in assessing and comparing their

systems. Among the most innovative and powerful mechanisms we suggest are ones involving assessing AGI systems' behaviors within the AGISim simulation world.

Assessing the ethicalness of an AGI's behavior and cognition is a matter even less studied. Our primary focus in this regard will be on the creation of "ethical behavior rubric" in the form of scenarios within the AGISim world. This sort of assessment does not provide any sort of absolute guarantee of an AGI system's safety or beneficialness, but nevertheless will allow a far more rigorous assessment than any approach now available. We consider it important that work in this area begins soon, so that "ethics testing" becomes accepted as a standard part of AGI R&D.

- **Recognizing Situational Entailment Challenge.** We plan to extend the "Recognizing Textual Entailment" challenge by defining a "Recognizing Situational Entailment" challenge, in which AI systems are challenged to answer simple English questions about "simulation world movies" that they are shown. The movies will be generated using the AGISim framework. An annual workshop to address this challenge may be organized as part of a recognized AI conference.
- **Development of a Suite of Benchmark Learning Tasks within AGISim.** Within the context of the AGISim world, we will develop a set of tasks on which any AGI system can be tested, e.g. playing tag, imitating behaviors, imitating structures built from blocks, etc. Having a consistent set of benchmark tasks for comparing different AGI approaches is important for coordination of progress in the field.
- **Development of a Suite of Benchmark Ethics Tests within AGISim.** Just as one can test intelligence through AGISim scenarios, one can also test ethics, by placing the AGI in situations where it must interact with other agents, assessing the ethical sensitivity of its behaviors. Testing within such scenarios should become a standard part of assessing the nature of any new AGI architecture.
- **Porting of Human IQ Tests to AGIs.** To what extent are human IQ tests overly human-centric? Can we create variants of the IQ tests administered to humans that are more appropriate for AIs? It may be that different variants must be created for different AIs, e.g. based on the nature of the AIs embodiment and sensory organs. Investigating the variation of IQ questions based on the nature of the intelligent system being tested, is one way to probe the core of intelligence.

System Design & Implementation

Research Area 7: AGI Design

This is arguably the most critical component of the path to AGI. As noted earlier, AGI design and engineering will be our central focus in Phase II. In Phase I, however, our work in this area will focus on the comparison and formalization of existing AGI designs. This is crucial, as it will lead to a better understanding of the strong and weak points in our present understanding of AGI, and form the foundation for creating new AGI designs, as well as analyzing and modifying existing AGI designs.

- **Systematic Comparison of AGI Designs.** A number of designs for AGI have been proposed, some in the public literature, with varying levels of detail. What are their major overlaps, major common strengths, and major common weaknesses? The first step toward resolving this may be to describe the various systems using a common vocabulary, such as the Mind Ontology Project.

Research Area 8: Cognitive Technologies

Our in-house R&D is founded, in part, on the premise that appropriate use of probability theory is likely to play an important role in the development of safe, beneficial AGI. With this in mind, the "cognitive technologies" aspect of our Phase I centers on the creation of several cognitive components utilizing probability theory to carry out operations important to any AGI.

Our research in this area will differ from most work on probabilistic AI due to our focus on generality of scope rather than highly specialized problem-solving. In order to reason probabilistically about real-world situations, including situations where ethical decisions must be made, powerful probabilistic reasoning tools will be needed, and tools different-in-kind than ones currently popular for narrow-AI applications.

- **Efficient Techniques for Managing Uncertainty in Large Dynamic Knowledge Bases.** Using Bayesian probability techniques is not sufficiently computationally efficient to be a pragmatic approach to AGI on present systems. Approximations are needed, which achieve efficiency without losing too much accuracy. A variety of approaches are possible here, and need to be fleshed out mathematically and computationally, and compared to each other. For example, work on loopy Bayes nets, imprecise and indefinite probabilities, and probabilistic logic networks is relevant here.
- **Probabilistic Evolutionary Program Learning.** One of the more powerful optimization techniques available is "probabilistic evolutionary learning," or Estimation of Distribution Algorithms (EDAs). Recent research by Moshe Looks has extended EDAs to automated program learning, but the state of the art only allows automated learning of relatively simple programs. Extension of this paradigm is

necessary, to allow learning of programs involving recursion and other complex programming constructs.

- **Probabilistic Inference Driven Self-Modification of Program Code.** Is it possible to write code that uses probabilistic reasoning to model its own behavior, and then modifies itself accordingly? Two proprietary AGI designs, Novamente and Self-Aware Systems, use aspects of this idea. However, there is no general theory covering this kind of algorithm, and many possible approaches may be viable.

New: SIAI Research Grant Program

SIAI has finalized the structure for the SIAI Research Grant Program, in collaboration with the Future of Humanity Institute at Oxford. The initial trial period of two years will begin once sufficient funding is secured for the program. Research grants will be awarded to Ph.D. students, postdocs, or senior researchers pursuing promising research projects in one of our focus areas:

- Foundations of Reflection
- Friendly AI
- Existential Risks
- Smarter and Wiser

We are looking to support novel research to foster scientific understanding in the four areas, all of which remain severely underfunded by traditional institutions. **Grants will be awarded initially for a period of three mo., six mo., or one year.** Grants for Ph.D. students will range from \$5,000 to \$30,000. Grants for postdoc and senior researchers will range from \$10,000 to \$80,000.

SIAI will have no preference to an applicant's project location, institutional affiliation, country of residence, or nationality. SIAI will review applications via a two-stage process: **1)** Applications will be screened against SIAI's Evaluation Criteria, and then notified whether they are successful at this stage, generally three months after the deadline. **2)** Successful applications will be sent to an external panel for academic review. Recommendations will then be made and presented to the SIAI Final Selection Board, which will determine by vote which applications are awarded.

Applications will be evaluated on the quality of the proposed research and its potential contribution to knowledge in one of our focus areas. SIAI will look for three characteristics:

- A well-defined research problem, objective, or question
- A sound assessment of what the research can achieve
- A clear explanation of the significance of the research

Foundations of Reflection Focus Area

In 1965, the famous statistician I. J. Good suggested that any sufficiently smart mind would be capable of designing a next generation of even smarter minds, giving rise to a positive feedback loop and an "intelligence explosion." In purest form, an AI could rewrite its source code, or design new hardware for itself. Taking this scenario seriously presents us with foundational mathematical challenges: modern logics, probability theories, and decision theories do not adequately handle this type of self-reference. These difficulties are symptomatic of a more fundamental problem – current AI techniques make little use of reflection. We, on the other hand, derive a great deal of benefit from thinking about thinking. What makes self-reference tractable for human reasoners? Can the same techniques be applied to AI and made reliable, and safe?

The Foundations of Reflection Focus Area will support research on the following: **1)** Decision-theoretic, probability-theoretic, or logical foundations for self-referential agents. **2)** Analyses of the stability of optimization targets, utility functions, and choice criteria in self-modifying agents. **3)** Techniques for self-modification and self-improvement in AI that promise to be strengthenable to extreme reliability. **4)** Formal attempts to prove the problem of ensuring extreme reliability in AI unsolvable, or unsolvable using a particular approach, using proof-theoretic or other mathematical arguments. **5)** Experiments to determine how human problem-solvers use reflection. **6)** Other work that demonstrates a significant new idea or approach.

Friendly AI Focus Area

The Foundations of Reflection Focus Area arises as part of our broader research focus on Friendly AI: handling the challenge and risk of smarter-than-human AI. SIAI supports technical work on this subject: for example, identifying the problems of learning a multicomponent utility function, under conditions of uncertainty, from noisy, non-independent observations.

The Friendly AI Focus Area will support research on the following: **1)** Conditions sufficient, or necessary, for an agent to learn information in a complexly structured utility function or other choice criterion. **2)** Acceptable ways of transforming human motivational structure (e.g. adaptation aspiration) into normative criteria such as utility functions. **3)** Detailed identification of known biases or logical fallacies in previously published work on AI motivations. **4)** Rigorous analytic philosophy that addresses some of the biggest challenges in Friendly AI, such as “How do we know what we want?” **5)** Clear explanations of how an existing AI methodology would fail in Friendliness when scaled to superintelligence, self-modification, real-world capabilities exceeding the programmers, metamoral questions, or other long-term challenges of Friendly AI.

Existential Risks Focus Area

Nearly 99.9% of all species that ever lived are now extinct. Will our own species have the same end? How could that happen? And what can we do to stave off the end? An "existential risk" is defined as one that threatens to annihilate Earth-originating intelligent life or permanently and drastically curtail its potential. Existential risks are the extreme end of global catastrophic risks. The exact probability of an existential disaster in this century is unknown, but there is reason to think that it is significant. That, at least, is the consensus among the small number of serious scholars who have investigated the question: **50%**, Professor Sir Martin Rees, President of the Royal Society (*Our Final Hour*, 2003); **30%**, Professor John Leslie (*End of the World*, 1996); **significant**, Judge Richard Posner (*Catastrophe*, 2004); and **not less than 25%**, Dr. Nick Bostrom ("Existential Risks: Analyzing Human Extinction Scenarios," 2002).

The Existential Risks Focus Area will support research on the following: **1)** Original studies and metaresearch on existential risks. **2)** Comparative analyses of the threats posed by existential risks. **3)** Methodological improvements for studying existential risks. **4)** Analyses of complex ethical issues related to existential risks (such as the weight to be placed on the interests of future generations, and how to allocate moral responsibility for risk-reduction). **5)** Identification of common elements that contribute to existential risks.

Smarter and Wiser Focus Area

The quality of human thinking is a bottleneck resource for humanity in the 21st century. From finding cures for diseases, to improving manufacturing and business processes, inventing new technologies, addressing environmental problems, solving scientific mysteries, anticipating and managing risks, and negotiating workable approaches to complex ethical and political issues: in these and many other areas, progress is critically dependent on the abilities of human individuals and collectives to gather, evaluate, filter, organize, and aggregate information. Even modest improvements of these abilities could have a significant enabling effect on society's capacity to meet challenges on many fronts in the coming years and decades. A better understanding of the means by which we can become smarter and wiser could pay enormous social dividends.

The Smarter and Wiser Focus Area will support research into how human epistemic capacities can be improved on three levels: **1)** Enhancement of individual cognitive capacities. **2)** Enhancement of the ways society or groups process information. **3)** Improvements of methodological and epistemological standards.

Under Development: The AI Impact Initiative

Advanced AI has the potential to impact every aspect of human life. We are in a critical window of opportunity where we have powerful but temporary leverage to influence the outcome. Only a small group of scientists are aware of the central issues, and it is essential to get input from a

broader range of thinkers. **The AI Impact Initiative** will foster an interdisciplinary framework for the safe and beneficial deployment of advanced AI. We will form a multidisciplinary body of experts to bring a broad perspective to the critical issue of advanced AI's impact on humanity. This effort will involve researchers with expertise in many different fields, including computer science, security, cryptography, economics, industrial organization, evolutionary biology, ethics, cognitive science, political theory, decision theory, physics, philosophy, religious thought, etc.

The AI Impact Initiative will host meetings over a period of three to five years to analyze the central issues and produce strategic recommendations. An inaugural workshop will be held in '08 or '09 to bring together researchers from a variety of disciplines and begin the process of unifying their insights. One of the goals is to produce documents that clearly express the central issues so that a broader group of participants may usefully contribute. The Initiative's longer term goal is to lay the foundation for a new science to study these issues. This will involve creating expository materials, building an international network of scientists and scholars, organizing workshops, and creating a comprehensive report to provide direction for future research and development.

SIAI in 2006: Publications and Presentations

Publications

2006 – Cognitive Biases Potentially Affecting Judgment of Global Risks

2006 – Artificial Intelligence as a Positive and Negative Factor in Global Risk

2006 – Knowability of Friendly AI (work in progress: <http://sl4.org/wiki/KnowabilityOfFAI>)

Presentations

2006 – AGI: What are the Risks? (2nd AGI Workshop)

2006 – Artificial Intelligence as a Precise Art (1st AGI Workshop)

2006 – The Human Importance of the Intelligence Explosion (Singularity Summit)

2006 – The World's Most Important Math Problem (Bay Area Future Salon)

SIAI in 2006: Research Update

In July, Herreshoff and Yudkowsky began working together on AI theory full-time, which has significantly sped up the rate of progress. Conversely, however, Yudkowsky has had less time to write publications and present at conferences. In July and August, Nick Hay (a Masters candidate at the University of Auckland) and Peter de Blanc (a math graduate from the University of Pennsylvania) worked with Herreshoff and Yudkowsky on AI for six weeks under an internship.

SIAI's 2006 research consisted of analyzing, in excruciating detail, many small “toy” problems that humans seem to solve using thinking about thinking, looking for a key insight that explains

why these problems are tractable. These toy problems could be easily solved by writing a specialized AI program; but SIAI is looking for the metaprogram that lets humans look over the problem and then write a program to solve it. Since metaprogramming is, in general, a difficult and unsolved problem, SIAI is analyzing it for simple cases to make the problem tractable. The research of Yudkowsky and Herreshoff has produced some interesting insights. SIAI is looking to hire a senior science writer who, as part of their role at SIAI, will begin to report a backlog of past results that includes mathematical definitions of optimization, criteria of intelligence within a Bayesian framework, and hidden downsides of randomized algorithms in artificial intelligence.

SIAI in 2006: The Singularity Summit at Stanford

The Singularity Summit at Stanford presentations can be viewed online or downloaded at SIAI's website. SIAI organized the event in partnership with the Symbolic Systems Program, Center for the Study of Language and Information, Stanford Transhumanist Association, and KurzweilAI.

Presenters included:

- **Dr. Nick Bostrom**, Director of the Future of Humanity Institute at Oxford University
- **Cory Doctorow**, Visiting Professor at USC Center on Public Diplomacy
- **Dr. K. Eric Drexler**, Chief Technical Advisor of Nanorex Inc.
- **Dr. Douglas R. Hofstadter**, College Professor of Cognitive Science and Computer Science at Indiana University Bloomington, Author of *Gödel, Escher, Bach*
- **Steve Jurvetson**, Managing Director of Draper Fisher Jurvetson
- **Bill McKibben**, Author of *The End of Nature* and *Enough*
- **Dr. Max More**, Futurist and Cofounder of Extropy Institute
- **Christine Peterson**, Cofounder and VP of Public Policy at Foresight Nanotech Institute
- **John Smart**, Founder and President of Acceleration Studies Foundation
- **Dr. Sebastian Thrun**, Director of Stanford AI Laboratory
- **Peter Thiel**, Founder and Managing Member of Clarium Capital Management
- **Eliezer Yudkowsky**, Cofounder and Research Fellow of the Singularity Institute

Select Blog and Press Coverage

San Francisco Chronicle: Smarter than thou?

<http://www.sfgate.com/cgi-bin/article.cgi?f=/c/a/2006/05/12/BUG9IIMG1V197.DTL>

ZDNet: The great Singularity debate

<http://blogs.zdnet.com/BTL/index.php?p=3029>

Another View of the *Singularity Summit*: Is the “Singularity” for Everyone?

http://www.thehumanfuture.org/commentaries/commentary_willow.html

Jerry Pournelle's Report on the Singularity Summit

<http://www.jerrypournelle.com/reports/jerryp/singularity.html>

Fight Aging: Roundup on the Singularity Summit at Stanford

<http://www.fightaging.org/archives/000847.php>

Down the Avenue: Singularity Summit Opens

http://www.downtheavenue.com/2006/05/singularity_sum_1.html

Center for Responsible Nanotechnology's Summit Coverage

http://crnano.typepad.com/crnblog/2006/05/singularity_sum.html

SIAI in 2006: SIAI Canada

SIAI-CA continued to fulfill its objective as the Canadian 'on-ramp' for supporters of SIAI's goals. As a designated charitable organization, SIAI-CA issued receipts to all its donors, who may now obtain tax relief for 2006. SIAI-CA continued to be run by unpaid volunteers, and nearly 100% of SIAI-CA donations went to supporting the organization's goals, not expenses.

2006 began with high donor activity in light of the 2006 Singularity Challenge, which ran until February. All challenge donations were matched by Peter Thiel. (Thank you to everyone who gave.) February also saw SIAI-CA as the subject of a positive editorial by Keith Norbury of Victoria News, a local newspaper, in Victoria, British Columbia, SIAI-CA's home town. See:

<http://lists.extropy.org/pipermail/extropy-chat/2006-February/024678.html>

The president of the Canada branch, Michael Roy Ames, continued to respond to "SIAI Feedback" emails – an avenue for the public to ask questions about SIAI and singularity issues. Ames also assisted Tyler Emerson with preparations for the 2006 Singularity Summit at Stanford.

In December, SIAI-CA awarded Marcello Herreshoff a grant of \$7,500. This will allow Herreshoff to work between January 2007 to August 2007 on reflection and Friendly AI theory. Also in December, SIAI-CA's webpages were integrated into SIAI's website revision. See:

<http://www.singinst.org/aboutus/siaicanada>

SIAI in 2006: Website Revision

As part of the 2006-2007 SIAI Strategic Marketing Plan, the SIAI website was updated and revised in November 2006. The new SIAI website can be seen at <http://www.singinst.org>.