DARPA-BAA-10-53

The Mind of the Mind's Eye

Technical Area 1: Visual Intelligence

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Computer Science and Artificial Intelligence Laboratory (CSAIL)

Technical Point of Contact:
Professor Patrick H. Winston
CSAIL
Massachusetts Institute of Technology
32 Vassar Street
Cambridge, MA 02139
Email: phw@csail.mit.edu
Phone: (617) 253-6754
Fax: (617) 258-8682

Administrative Point of Contact:
Laureen E. Horton
Office Sponsored Programs, E19-750
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139
Email: laureena@mit.edu
Phone: (617) 253-3922
Fax: (617) 253-4734

Summary of Costs:

Base Y1: 288,445	Y2: 298,049	Y3: 303,037	Total: 889,531
Option 1	Y4: 303,424	Y5: 314,227	Total: 617,651

Business Type: Other Educational

Laureen E. Horton
Manager, Contract & Grants

Massachusetts Institute of Technology
77 Massachusetts Avenue, Building E19-750
Cambridge, Massachusetts 02139-4307

Office of Sponsored Programs

Phone 617.253.3922
Fax 617.253.4734
Email laureena@mit.edu
http://web.mit.edu/osp/www/

May 7, 2010

James Donlan
Program Manager, DARPA-BAA-10-53
DARPA/TCTO
3701 North Fairfax Drive
Arlington, VA 22203-1714

RE: DARPA-BAA-10-53

Dear Dr. Donlan:

The Massachusetts Institute of Technology (MIT) submits herewith a new proposal entitled, "The Mind of the Mind's Eye" for the above-referenced BAA. The research will be performed under the direction of Patrick Winston as Principal Investigator in MIT's Computer Science and Artificial Intelligence Laboratory.

We are requesting funding in the amount of $1,490,432 in total estimated costs for the period of September 1, 2010 through August 31, 2015.

The research program to be conducted by MIT is considered contracted fundamental research consistent with DOD Instruction 5230.27, paragraph 4.3, research results are published and or disseminated without restriction.

We hereby affirm to the best of our knowledge and belief that no Organizational Conflict of Interest (OCI) exists in regard to the effort proposed under DARPA-BAA-10-53. Massachusetts Institute of Technology does not have existing contracts with DARPA which provide Scientific, Engineering and Technical Assistance (SETA) support to any DARPA technical office.

Please understand that any acceptance of a resulting award for this program is contingent upon acceptable negotiations that result in a funding instrument that incorporates terms and conditions appropriate for a non-profit educational institution. MIT's policies prohibit the acceptance of awards that include provisions that restricts our ability to disseminate research results or place limitations on the use of foreign nationals working on research. MIT must be able to publish research results without restriction and must retain title to all intellectual property (with a non-exclusive license to the government).

Please direct any questions relating to technical aspects of the proposal to the Principal Investigator. Questions of an administrative nature may be directed to the undersigned at (617) 253-3922.

Sincerely,

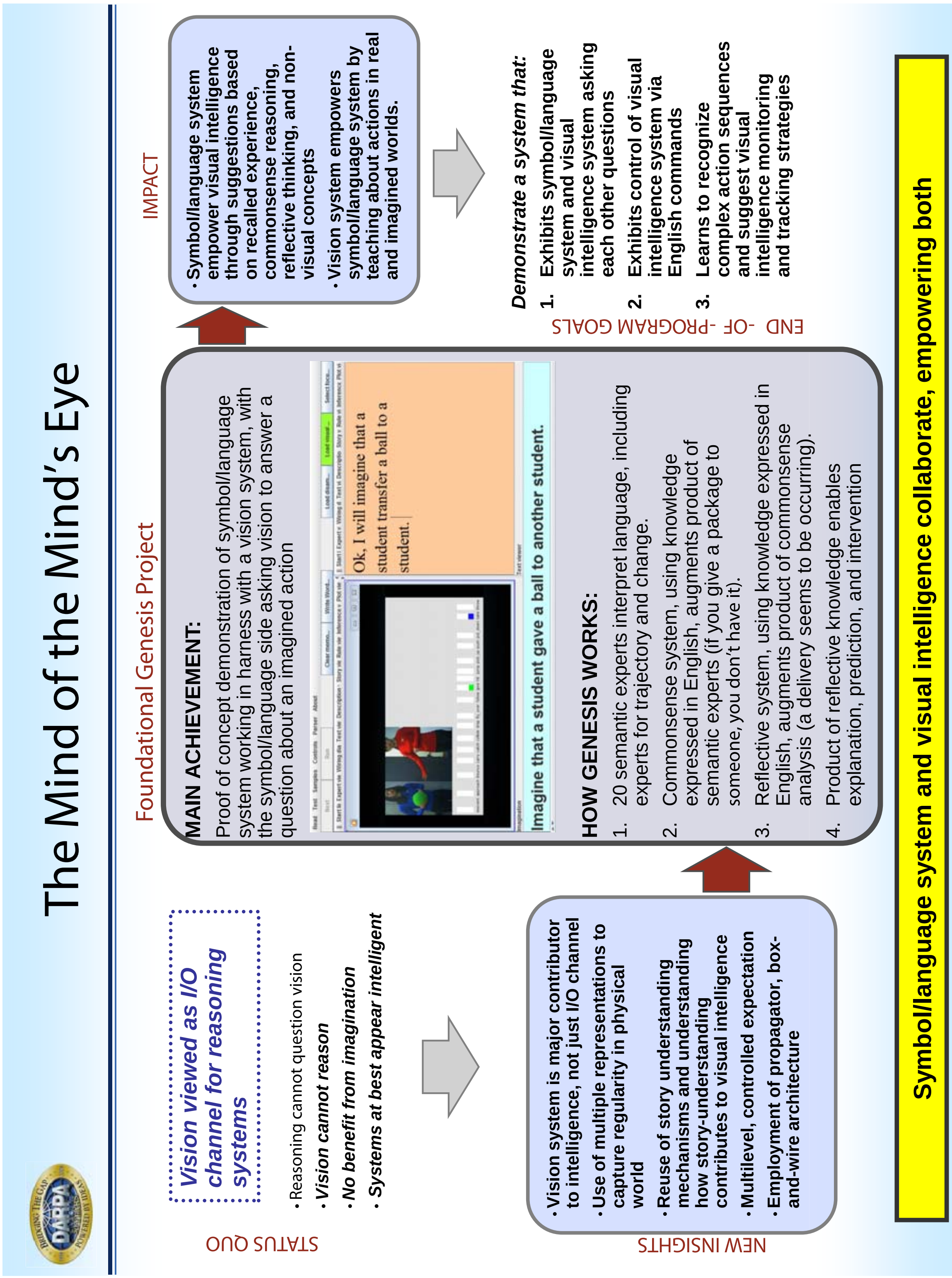
Laureen Horton

Laureen E. Horton
Manager, Contracts & Grants

1.3 Table of Contents

Section 2: Technical Details

2.1 Summary Chart



2.2 Innovative claims

The research proposed here, under the direction of Professors Patrick Winston, complements other research on visual intelligence, especially that separately proposed by Co57 Systems, under the direction of Dr. Sajit Rao.

The MIT CSAIL work will focus at the upper end on grounding, envisionment, and event-to-event expectation.

At the highest level, the innovative scientific idea behind the MIT CSAIL proposal is that any explanation of humanlevel intelligence must address the contributions of vision, language, and eventually the motor system, and deep interactions among them all. Since Turing's famous paper of 1950, most workers in Artificial Intelligence have viewed vision, motor systems, and language as the I/O channels for symbolic reasoning systems. We take a different view: we believe intelligence lies in the I/O systems, not behind them. Our view is explained and defended in detail in the white paper, *Taking Machine Intelligence to the Next, Much Higher Level, included in Section 3 of this volume.*

In the vision system, humans have dedicated representations and processes for the recognition, analysis, and imagination of events. In the language system, humans have representations and processes that enable the symbolic description and re-description of events, that construct and exploit regularities in those events, and that make generalizations.

We have used our *Genesis System* to refine and develop innovations that we propose to carry into the Mind's Eye Program in support of the Mind's Eye research goals. These innovations include the following:

- The use of many representations for physical events

Representations provide the means to capture regularities and build models. We believe language suggests representations that are particularly important for understanding the physical world: that which is easy to say is that which is important to be able to convey quickly and accurately. Accordingly, we have incorporated into the Genesis system representations for, for example, trajectory, location, place, transition (appear, disappear, increase, decrease), transfer, size comparison, coercion, cause, goal, belief, and mood. All these will be strengthened and augmented as we move forward in Mind's Eye research.

- The reuse of story-understanding apparatus

Event sequences populate our thinking. They go by various names—such as scenarios, stories, narratives, and cases—depending on context. No matter what the context, however, we believe they are understood on many levels, be they like the story of the woman entering and leaving a building, told in the BAA, or a complex military operation, or the greed-and-murder scenario with which we have worked out many details of story understanding. On one level, there is that which is explicitly observed (the woman gives a package to a man, Macbeth murders Duncan); that which is directly inferrible (the man has the package, Duncan is dead); and most importantly, that which is determined by reflective thinking about extended patterns in the events that constitute the story (the package transfer was involved in IED construction, the murder initiated a revenge pattern). All our inference and reflection mechanisms will be strengthened and augment as we move forward in Mind's Eye research.

- The idea of multilevel, controlled expectation

We have expectations at all levels in the seeing process. We expect what will be seen in the next video frame, at the completion of a give action, where the disappearing woman may emerge, and what will happen once the IED is completed. All this expectation dramatically reduces computation: we need not explore all possibilities; we mostly have to determine that new information is consistent with expectations. Thus, vision can be said to be a form of controlled hallucination. We have found that story understanding has the same quality. We infer many expectations from sparse information, and those expectations are pruned and guided as the story unfolds. Just as the possibilities seem about to explode, we ask ourselves a perception question meant to get us on the right track.

- The propagator architecture

Early attempts to build the Genesis system were soon blocked by complexity. Students were not able to learn enough about the system to make a contribution during a student lifetime. Fortunately, we were inspired by the work of Professor Gerald Sussman and his students, and we adopted their wired-box, propagator architecture. Each module is viewed as a box with ports and only interaction via the ports is permitted. A box is free to ignore information presented on a port if it is unexpected or unintelligible. New boxes are readily substituted in for old. No wild and distant system calls are permitted. All box interaction is transparent and visualized in a dynamic dashboard display. Wires can extend over the Internet, and indeed, the START parser used in our system is viewed as just another box even though it is physically at a location unknown to us. We propose to use this same architecture as we move forward, and in particular, our interface to system modules, such as those developed by Co57 Systems, will be via virtual, Internet-carried wires.

2.3 Roadmap

Our research **vision** is captured in the *Innovative Claims section* (page 7).

The **steps** we propose to take are enumerated in the *Workstatement* (page 22). The steps will be executed in the context of our guiding principles and existing software base, both described in detail in the *Technical Approach section* (page 12).

Some recent **news** is described in the *What Genesis does* subsection of the *Technical Approach section* (page 14).

Our expected **contributions** to the Mind's Eye program in general are treated throughout. Our contributions to persistent stare in particular are described in subsection *b* below.

a. Main goals of the proposed research

Our main goal is to provide the Mind's Eye Program with answers to the *grounding* problems discussed in the proposal and other grounding problems that are bound to emerge in the course of program execution. Our grounding effort will lead to a symbol/language system that supports visual intelligence by capturing the stories told by visual intelligence systems, by storing and recalling visual event sequences to support future visual intelligence, by learning characteristics of visual event sequences at the symbolic level, and by answering questions posed by visual intelligence systems. An additional goal is to provide an English command and question interface to visual intelligence systems, along with dashboard viewers to aid in debugging and analyzing experimental results.

b. Tangible benefits to persistent stare applications

Our special contribution to the Mind's Eye program lies in our proposed work on **grounding**, listed along with visual event learning, spatiotemporal patterns, environment, and visual inspection as the five areas of contribution in the BAA. Our determination to make this part of the Mind's Eye program successful reminds us of Joe Namath's comment just before Superbowl III: We guarantee it.

c. Significance of the approach to machine intelligence

We are committed to the idea that any explanation of humanlevel intelligence must address the contributions of vision, language, and eventually the motor system, and deep interactions among them all. Our commitment took shape in the BICA program and led eventually to our white paper, *Taking Machine Intelligence to the Next, Much Higher Level*, included in Section 3 of this volume (page 36). In light of our commitment, we see our work on this program as lying directly on the critical path toward understanding human intelligence in general, with extraordinary implications for applications of machine intelligence.

d. Critical technical barriers

Progress on computer vision has been retarded by viewing vision as an I/O channel. We deal with that barrier by putting vision and symbols/language on the same plane, viewing both as essential to intelligence, both as cooperating problem solvers, both as systems that can ask the other questions through wide communication channels.

Progress on symbol systems has been retarded by viewing symbol systems as problem solving apparatus not particularly requiring assistance from visual intelligence, and imagination. We deal with that barrier by appreciating the problem-solving strength of visual intelligence and especially visual intelligence backed by imagination.

Progress on language systems has been retarded by focusing language system research on statistical syntax analysis. We deal with that barrier by focusing on what language systems can do for systems aimed at instantiating representations that capture regularities in what perception delivers about what is going on in the physical world.

e. Main elements of the proposed technical approach

Our technical approach builds on the principles embodied and the software embedded in the Genesis system. Our approach will enable a balanced partnership between symbols/language and visual intelligence.

We will be guided by fundamental principles, articulated in the *Technical Approach Section* (page 12), that collectively constitute the core of our technical approach. The principles include:

The grounding principle. Our symbol/language work will be aimed at tight coupling with visual intelligence from the START, rather than supposing that symbol/language systems should interact with visual intelligence systems like the interaction of the top and bottom of an hour glass. **The Rumpelstiltskin principle**. Visual Intelligence has been neglected. At the same time, higher level intelligence surely involves symbols and language, because without them, thoughts cannot be indexed, concepts cannot be combined, and analogical reasoning is severely limited. We call it the Rumpelstiltskin principle because symbols (also known as names) give us power over ideas. **The multimodal principle**. Taken together, the grounding principle and the Rumpelstiltskin principle tell us that language and perception are both essential to intelligence. **The representation and multiple representation principle**. Models enable understanding, because models express constraint and regularity and constraint and regularity enable explanation of the past, prediction of the future, and control via intervention. Hence, we need representation, and more than one representation, because there is more than one kind of constraint and regularity to be captured. **The layering principle**. An intelligent system centered on the Mind's Eye will make use of representations of constraint and regularity on many levels. Thus, the systems of a Mind's Eye system will interact up and down (from low-level camera frames to high-level visual events) as well as sideways (vision working with language). **The Goldilocks principle**. The intermediate-sized feature principle seems at work in dealing with event sequences as well as in visual recognition. The best precedents are found not on the basis of the characteristics of low level features, because that filter is too weak, and not on the basis of good high-level match, because no event sequence is exactly like another. Hence, we will use intermediate characterizations, such as digging a hole, in our Mind's Eye work. **The leapfrogging principle**. We will use existing systems insofar as possible. In our work to date, we have used Wordnet as a source of classification information and Boris Katz's START parser. **The propagator architecture**. Many ambitious attempts to build complex systems collapse under the weight of their own complexity. We use the recently developed propagator architecture, which offers a fresh approach that takes us beyond ordinary modularity.

We will build on the Genesis system, a system that embodies our principles (page 12). The Genesis system has interacted with a visual intelligence system and has served as the foundation for a story understanding system. Both capabilities will help to jumpstart the Mind's Eye program.

The Genesis system makes use of approximately 20 representations (page 15). Of these, the most important address actions that can be sensed in the physical world. Of these representations, trajectory and transition are conspicuously common, not only in discussing what can be seen but also movement and change in abstract worlds. **All knowledge in the Genesis system is expressed in English.** Thus, knowledge is easy to add, readily understood, exposed for debate, and readily reused by any other system capable of absorbing knowledge expressed in English. **Genesis representations are built from just four Java classes.** Accordingly, it is easy to communicate with systems separate from Genesis, written in any programming language. **The Genesis system consists of approximately 150 boxes connected together with wires** (page 18). The architecture enables implementers to focus on individual boxes, eliminating the need to understand the system as a whole. **Genesis makes use of START and Wordnet** (page 18). Neither is perfectly suited to our objectives, both help us get to the central research issues immediately.

f. Basis of confidence

We are confident in success for several reasons: First, we have participated centrally in the construction of a precursor system. Second, we have demonstrated commitment to a world view that is congruent with the goals of the program. Third, We see the work as incredibly exciting from both the scientific and applications points of view.

g. Nature and description of end results to be delivered to DARPA

We will deliver the following, described in detail in the *Statement of Work* section (page 22):

- Source code for a system, built on a Genesis base, specialized to the needs of the Mind's eye program, starting at or before the end of the first year and continuously updated thereafter.
- A Webstart version of the system, starting at or before the end of the first year and continuously updated thereafter.
- An evolving specification describing the representations used in the Mind's Eye interactions with vision systems, including new representations and those already in place.
- Journal papers, conference presentations, and theses reporting on experimental results.
- Presentations at DARPA-sponsored program meetings.
- A final report, to include experimental results, a system description, final source code, a final representation specification, and a bibliography of papers, presentations, and theses.

h. Cost and schedule of the proposed effort.

We provide a project schedule in the *Schedule and Milestones* section (page 31). The total cost, with year 4 and 5 option, is $1,507,182.

2.4 Technical Approach

Our technical approach is straightforward: we will develop ideas and construct a system that builds on the principles embodied and the software embedded in the Genesis system. Accordingly, we begin with a statement of what Genesis does; then we enumerate the Genesis principles we propose to exploit; finally, we describe the Genesis software we propose to use to jumpstart our work.

Our proposed system will ensure a balanced partnership

Our central goal is to contribute the symbol/language side of a balanced partnership between visual and symbolic representations with both sides emphasizing capture of physical world constraints and visual exploitation of those constraints. Both sides will engage the other, as in these examples:

Language engages vision: A symbol/language system receives information about a situation that it has never encountered before: "The terrorists retreated into a building; prevent their escape." The symbol/language system determines that it must ask the visual intelligence system to monitor the doors and lower windows on all sides.

Vision engages Language: A visual intelligence system looks at the first and last parts of a event, such as an event that begins with a man with a rifle talking with a woman and ends with just a man and no rifle. The visual intelligence system determines that it must ask the symbol/language system to invent a likely story and conclusion, based on previous experience, such as "The man gave the rifle to the woman."

We will deploy enabling principles

To make such cooperative problem solving possible, we propose to be guided by fundamental principles that collectively constitute the core of our technical approach.

The grounding principle

Purely symbolic systems, without grounding in perception, cannot escape from what amounts to symbolic tail chasing and remain forever limited to knowledge that has been hard coded or explicitly told. Even CYC and Open Mind Commonsense are limited because there will always be facts that no one has thought to express. Accordingly, grounding in perception is essential. Without grounding, system builders can only hope to build systems that appear intelligent, but lack the depth that grounding provides, thus subjecting them to grotesque commonsense blunders.

The Rumpelstiltskin principle

Purely perceptual systems, like nonhuman primates, cannot reach the intelligence of people because without symbols and language, thoughts cannot be indexed, concepts cannot be combined, and analogical reasoning is severely limited. Symbols, in the end, give us power over ideas. Without symbols and language, system builders can only hope to build systems that react to the world without reflection on goals, question asking, and previous experience reuse.

The multimodal principle

Taken together, the grounding principle and the Rumpelstiltskin principle tell us that language and perception are both essential to intelligence. Each produces easy answers to questions the other would find difficult or impossible.

The representation and multiple representation principle

Models enable understanding, because models express constraint and regularity and constraint and regularity enable explanation of the past and prediction of the future.

Representations are the structures in which constraint and regularities are expressed, so without representations, there can be no models, no explanations, no predictions, and no interventions. Because there are many kinds of constraint and regularity, there must be many representations, each capable of exposing a constraint and enabling computations with that constraint.

It follows that the mind of the Mind's Eye will require multiple representations expressing constraints and regularities of many kinds, from statistical regularity to the more concrete regularity that emerges from the Newtonian physics of the world and the social interactions of people.

The layering principle

An intelligent system centered on the Mind's Eye will make use of representations of constraint and regularity on many levels. Thus, the systems of a Mind's Eye system will interact up and down (from low-level camera frames to high-level visual events) as well as sideways (vision working with language).

The Goldilocks principle

Ullman showed how to find faces using intermediate-sized features. You cannot look for eyes, because you will see one in every doorknob, and you cannot look for whole faces, because no whole face correlates well with other faces. Instead, you need libraries of intermediate-sized features—a nose and a mouth, two eyes and a forehead.

We have found the same intermediate-sized feature principle at work in dealing with stories. The best precedents are found not on the basis of the characteristics of the agents and objects, because that filter is too weak, and not on the basis of a literal resemblance to a precedent, because no story is exactly like another. Hence, we use intermediate features in our story work, and propose to use intermediate-sized features, in our Mind's Eye work. For example, we will take "Burying an object" to be an intermediate feature. It is not too small, because it consists of a several actions. It is not too big, because is not a whole story with motives and consequences.

The leapfrogging principle

In our work to date, we use Wordnet as a source of classification information and Boris Katz's START parser to perform syntactic analysis. Neither is perfect; both are good enough to get us to the research problems not yet tackled.

The propagator principle

Many ambitious attempts to build complex systems collapse under the weight of their own complexity. Examples readily come to mind from AI research and combat-system development. Fortunately, we have found that the propagator architecture offers a fresh approach from which we will benefit and to which we expect to contribute as a side effect of our work.

We will build on Genesis success

Our Genesis system adheres to the principles listed. Our plan is to take what we have built to another level, addressing the challenges raised when working with the visual intelligence part of the Mind's Eye Program. The Work Statement explains exactly the steps involved in exploiting the principles and taking what we have built to another level in another direction. Here we provide a brief overview of the Genesis system.

What Genesis does

The Genesis system is the anvil on which we have hammered out ideas in support of our view that symbol/language systems make the human species different because symbol/language systems enable our perceptual systems to do more than those of a chimpanzee and because symbol/language enables the understanding, recording, and reuse of event sequences.

On the perceptual side, Genesis has demonstrated how a language/symbol system can ask a vision system to imagine an event, ask the vision system to answer question using the imagined event, report the answer back to the symbol/language system, with the symbol/language system then caching the result in symbolic memory, thus learning something definite about the physical world.

In the defining experiment, we asked Genesis to ask a vision system to answer a question by imagining an action involving a transfer: **Imagine that a student gave a ball to another student.** The we asked **Did the other student take the ball?**

The vision system solved the problem with imagination, using visual routines that read the answer off of a stored, then-recalled scene, as shown in figure 1:

The vision system recalled the scene because, when analyzed visually, a *give* action was noted by a visual intelligence system that previously learned to recognize give actions. Then, the visual intelligence system answers the *take* question by noting that a *take* action occurred at the same time.

On the event-sequences side, Genesis is a story understanding, recording, and reuse system, because event sequences are stories. The power of Genesis on the story side has many exciting implications, including implications for the Mind's Eye program, because visual intelligence supplies information needed to tell stories about what is happening in the visual world. With a handle on an emerging story, it becomes possible to:

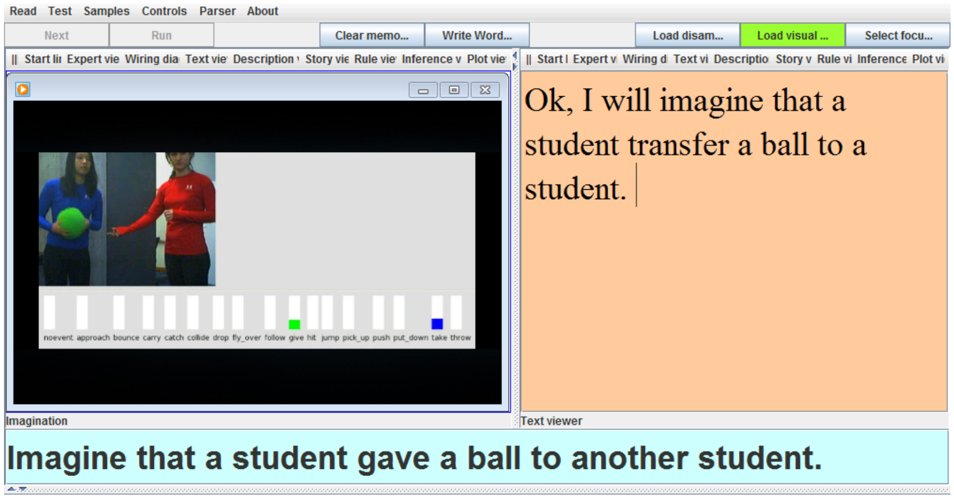


Figure 1. The Genesis system, working with a vision system, recalls a situation in which one student gives a ball to another. Because the vision system sees a *take* in the same sequence, Genesis notes that *give* and *take* co-occur.

- Relate the current scenario to other similar scenarios
- Suggest where to look to gather reinforcing or disconfirming information
- Exploit experience to determine what might happen next
- Suggest what ought to be monitored as a scenario continues to unfold

Today, Genesis regularly demonstrates that it understands complex event sequences, enriches information specifically provided with commonsense augmentations, reflects on what is happening, and suggests what might happen next.

Research-driving stories include the well-known Russian Cyber attack on Estonia's national network, civil war battles, a few legal cases and cyber-attack scenarios, and plots drawn from the work of Shakespeare. Most of our testing and debugging has been done on Shakespeare's *Macbeth* because the plot is familiar and because it is full of the greed, desire, emotion, and violence that characterize conflict in general, be it between people or nations.

Our most recent exercise of the Genesis system produces multiple interpretations of the same story, as if the story were viewed from the perspective of multiple cultures, with each interpretation colored by nuanced differences in commonsense and reflective knowledge.

In figure 2, for example, Macduff's killing of Macbeth is seen on one side as an act of insanity and on the other, as part of a revenge scenario.

Enabling Representations

The Genesis system is rich in representations because we believe various sorts representations enable systems to build models that capture various sorts of constraint and various sorts of regularity. Our representation set improves and expands as experiments indicate what needs to be strengthened and added.

As indicated in figure 3, Genesis makes use of approximately 20 representations, six of which—trajectory, location, place, transition, and comparative size—are conspicuously oriented toward what can be sensed in the physical world. Of these, two—trajectory and transition—are conspicuously common, not only in discussing what can be seen but also movement and change in abstract worlds.

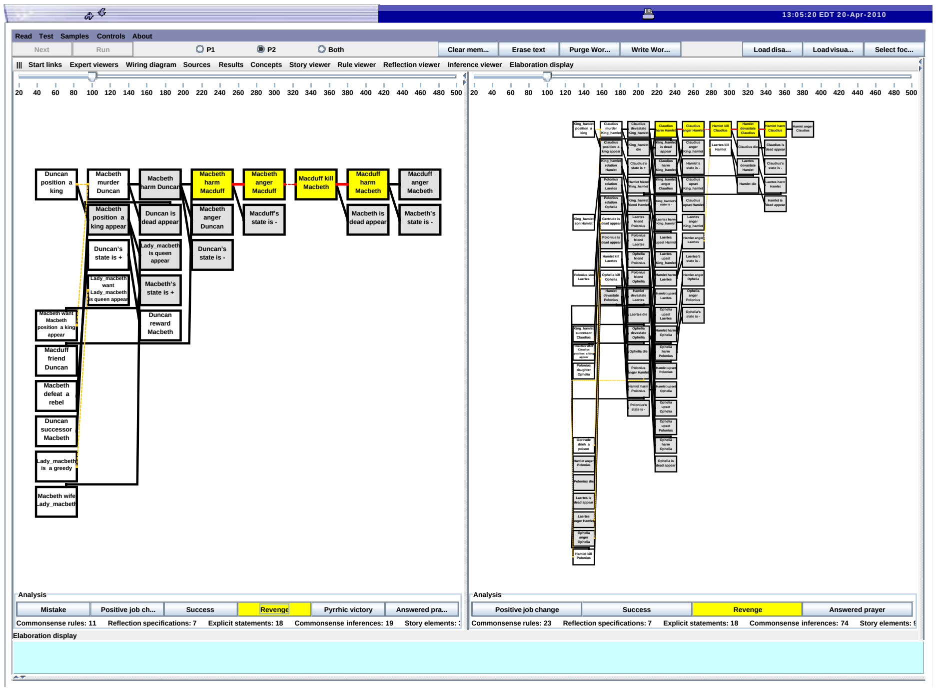


Figure 2. The Genesis system interprets a story in two culturally sensitive ways. On the left, a killing is seen from the perspective of cultures that view killing uniformly as acts of insanity. On the right, the same killing initiates a search for a reason, leading to a conclusion that revenge is involved.

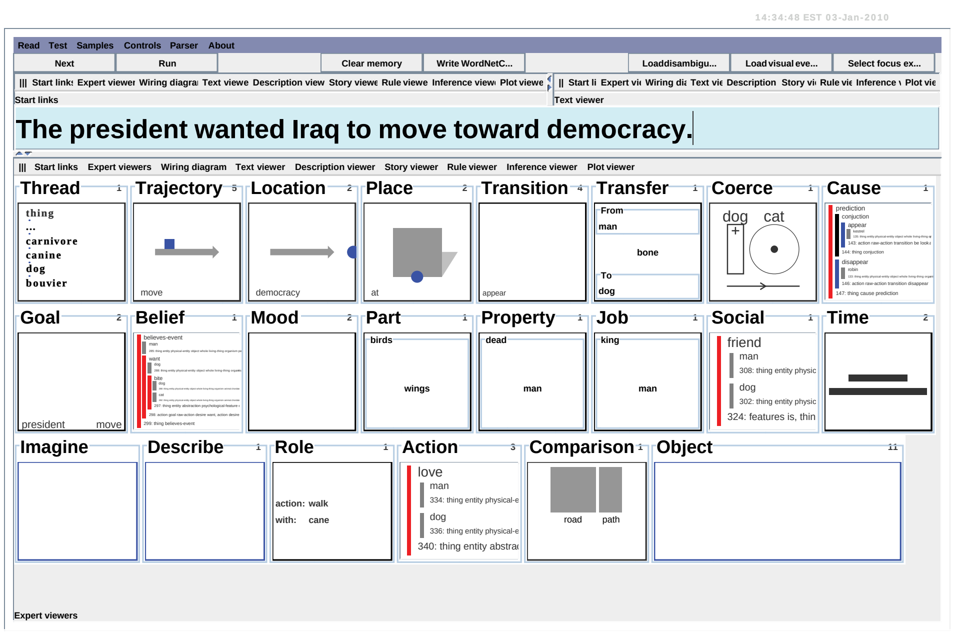


Figure 3. The Genesis system interprets a story in terms of multiple representations, the most important of which deal with movement and change in physical and abstract worlds. The snapshot reflects the results obtained from a test suite that includes sentences about dogs, kings, and Iraq moving toward democracy.

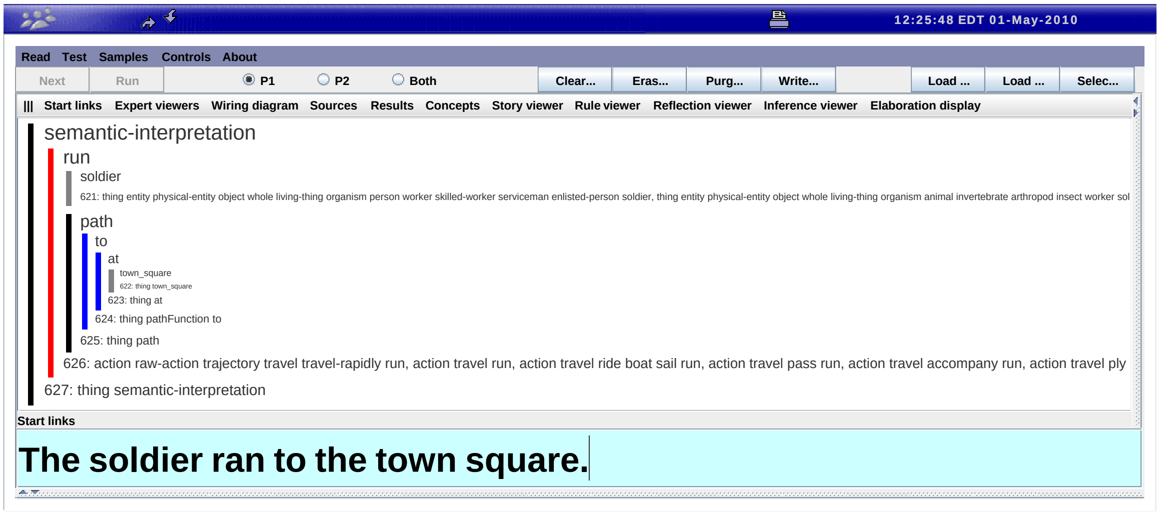


Figure 4. Representations in the Genesis system are implemented in terms of sequences, relations, single-sided relations, and plain objects. The *soldier* is a plain object belonging to many classes, including serviceman. *Run* is a relation between the soldier and a path. A *path* is a sequence because there can be any number of places on a path. *To* is a single-sided relations that indicates where the soldier ends up.

Knowledge expressed in English

All interpretation knowledge used in the Genesis system is exposed and transparent because all that knowledge is expressed in English. For example, here is a commonsense knowledge example, exactly as it is provided to Genesis:

```
If someone kills you, then you become dead.
```

And here is a reflective knowledge example, exactly as it is provided to Genesis:

```
Start description of "revenge".
xx is an entity.
yy is a entity.
xx's harming yy leads to yy's harming xx.
The end.
```

Because all knowledge is expressed in English, all knowledge is easy to add, readily understood, and exposed for debate. Also, the knowledge is an asset that can be used by any other system capable of absorbing knowledge expressed in English.

Grounding in four classes

All our representations are implemented on top of a small set of Java classes, as illustrated in figure 4. One class expresses a sequence of objects, another expresses relations between objects, yet another is a single-sided relation with just one object, and finally, there are plain objects, lacking any finer structure other than information about the types to which they belong.

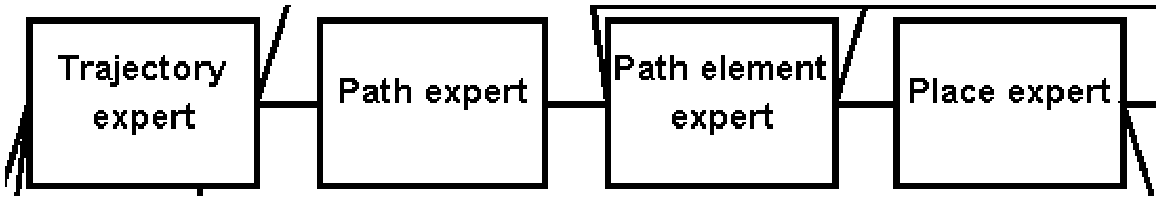


Figure 5. Genesis exploits ideas from Sussman's propagator architecture. The modules are boxes connected by wires.

Because all representations ground out in just four classes, it is easy to implement systems separate from Genesis, in any programming language, that can work with descriptions provided through Genesis representations.

Propagator boxes and wires

The Genesis system consists of approximately 150 boxes connected together with wires like those shown in figure 5. Some of the boxes are simple viewers. Two encapsulate large systems developed over many years by other researchers. The architecture enables implementers to focus on individual boxes, eliminating the need to understand the system as a whole. All an implementer needs to understand is the classes out of which representations are built and the outputs to be provided on output ports in response to signals arriving on input ports.

Simple function calls put in place the connections such as those shown in the figure:

```
Connections.wire(  getTrajectoryExpert(),   getPathExpert());
Connections.wire(  getPathExpert(),         getPathElementExpert());
Connections.wire(  getPathElementExpert(),  getPlaceExpert());
```

Major subsystems: START and Wordnet

Because Genesis is not a natural language research system, but needs to interpret ideas expressed in English, we use Boris Katz's START parser to perform syntactic analysis and we use Wornet to provide type information. Both are encapsulated in boxes, hence both are readily replaceable.

In early versions of Genesis, we used the off-the-shelf Stanford statistical parser. Like all statistical parses, it parsed everything; alas, like all known statistical parsers, it was unreliable and erratic. Changing a single word, such as **soldier** to **terrorist**, often produced a completely different statistical interpretation, making it extremely difficult for us to write systems to translate the statistical parse into semantic understanding, and even more difficult to write systems that learned to do the translation.

Accordingly, we switched about one year ago to Katz's START parser. Unlike the statistical parsers, it does not always produce a syntactic interpretation, but when it does, the result is reliable and in the form of a semantic net, which is much closer to the semantic understanding than a syntactic parse tree.

START was built as a question answering system, not a system to deal with statements and commands. Nevertheless, easy adjustments to START, encouraged by our work, have extended START so as to deal with the basic statements and commands we need, and as we need more, START grows in proportion. Katz's group has added, for example, if–then statements, statements with variables, statements with *occur* constructions, and many more, to support our work:

```
If someone kills you, then you become dead.
xx's performing an action leads to xx's becoming unhappy.
XX becomes happy because XX wanted an action to occur and the action occurred.
```

We are not completely happy with Wordnet, from which we extract type information. Wordnet often returns too many interpretations, but nevertheless often fails to return the expected interpretation as its first interpretation or even at all. Wordnet provides 15 entity interpretations for *Run* along with 36 action interpretations. Likewise, we have to tell our system that *John* is the name of a person, rather than a label for vulgar concepts.

Fortunately, we have the option of starting over, providing our own classification information via English sentences, such as “A Bouvier is a kind of dog.” We will do this to the extent needed.

2.5 Prior Work

The principal investigator, Professor Patrick Winston, has worked in close collaboration with Dr. Sajit Rao on a DARPA seedling program that put in place proofs of concept in support of the development of DARPA Mind's Eye Program. The research proposed here takes that collaboration to another level and initiates collaborations with other Mind's Eye contractors.

2.6 Comparison with Current Technology

First, vision systems are built in isolation and viewed primarily as systems for recognition and change-detection. Our research sees vision in a different light: as a place where thinking takes place.

Second, vision systems are viewed as input systems for reasoning systems that lie behind them. Information flows into the symbol world as if into a black hole, rarely and minimally communicating back to perception. Our research sees vision and language as constantly interacting, with each asking the other for essential answers to essential questions.

Third, complex systems today are modular and built in accordance with model-view-controller architecture, but big systems remain fragile, hard to modify, and hard to understand. We bring the propagator architecture to the Mind's Eye program specifically so as to fight off the big-system problems that are bound to emerge.

2.7 Statement of Work

The Mind of the Mind's Eye

In pursuing our goals, we will of course honor the principles explained in the Technical Approach section and exploit that which we have already incorporated into our Genesis system.

Because we are breaking new ground, the development of precision and recall evaluation metrics are is itself a research task. Until we develop such metrics, the fundamental test of success will be that visual imagination systems demonstrate capability if they use our contributions but not otherwise. Early demonstrations will be considered successful if results are deemed reasonable by developers and other program participants.

We are especially confident that we will be able to make rapid and on-schedule progress during the first three years because our experience with the Genesis system gives us a good sense of the work that needs to be done and confidence that it can be done without the emergence of any show-stopping blockers. As we go further into the program, of course, we expect that the most difficult problems we solve will be problems that have not yet appeared. Thus, we expect steady, predictable progress on the first four tasks and early work on Task five. Later work on Task Five and work on Task Six—being further in the future and less well defined—involves more risk.

Task 1: Develop robust, always-on pipes between Genesis and vision systems

The propagator box-and-wire architecture of the Genesis system has proved itself convincingly in Genesis development, enabling robust behavior, fault isolation consequences, easy division of system development tasks, straightforward module replacement, rapid progress up the system learning curve, and network access to major subsystems.

We need to extend the propagator box-and-wire architecture by completing the existing Java implementation and adding a C++ implementation and a MATLAB implementation. We also need to develop a box-and-wire based bidirectional message specification for communication between and among language and vision systems operating at multiple network sites.

Our approach will be to specify, develop, and test.

Milestones for Task 1

- October 2010:
  Specify propagator Internet and programming language connection needs for Java, C++, and MATLAB.
  Deliver specification document.
- February 2011:
  Complete Version I connection software.
  Demonstrate capability, report on capability, and deliver software.

Task 2: Take command coverage to next level

In preliminary work, we have readily handled sentences such as "Imagine that a student gave a ball to another student," taking such sentences through syntactic analysis using the START parser and through our own semantic analysis, which leads to the conclusion that there is a transfer in which a ball goes along a path from the first student to the second student and that contact appears between the ball and the second student. We aim to take this capability to the next level by first determining what commands are most useful, then generalizing the START parser if necessary, and finally adding the semantic apparatus needed to translate those most-useful commands into our physical-world representation suite. This will enable Genesis to issue requests and questions to the vision system such as:

- Look for a man who gives a package to another man.
- Look for a tall man carrying a brown bag.
- Look for people appearing on the rooftops.
- Report anyone moving rapidly.
- Watch for someone carrying a shovel.
- Monitor the doors and count the people passing through.
- Focus your attention on the blue car.
- What is the man sitting on the bench doing?
- What is the tall man doing?
- Where is the woman going?
- Imagine [variations on the above].

Our approach will be to specify, develop, and test.

Milestones for Task 2

- February, 2011:

  Specify 10-20 basic English commands needed to direct visual intelligence.
  Deliver command specification.
- September 2011:

  Complete Version I English command software for basic commands.
  Demonstrate capability, report on capability, and deliver software.

Task 3: Enrich our representation suite

In preliminary work, we have exploited the trajectory representation (path, path element, place, etc.) developed by Ray Jackendoff, a linguist, and the transition representation (appearance of contact, increase in speed, etc.) developed by Gary Borschardt, a member of our team. Both of these representations were developed in light of the great frequency of trajectory and transition evoking words in English usage. Every 100 sentences in the well-known Wall-Street-Journal corpus contains 25 trajectories and transitions.

These representations will provide a foundation for moving forward because, in combination with each other and with our cause and transfer representations, they capture the high-level essence of most of the verbs of interest listed in the BAA. We will have to augment what we have, however, because some of the verbs involve aspects of physical movement not yet covered, such as specific direction or change in direction (turn, raise), repetition (bounce), and result (attach, chase). We further anticipate a need to handle adjectives of manner (hurriedly, erratically, purposefully).

Our approach will be to build upon the use of representational diversity in the Genesis system, taking the existing software base, representations, and concepts already demonstrated to the next level. See note on evaluation at the beginning of this Work Statement.

Milestones for Task 3

- February 2011:
  Identify low-level representations driven by English-command needs in support of Task 2.
  Report on representational needs.
- September 2011:
  Complete implementation of Version I low-level representations in support of Task 2.
  Demonstrate capability, report on capability, and deliver software.
- February 2012:
  Identify representations needed for learning action consequences in support of Task 4.
  Deliver specification.
- September 2012:
  Complete implementation of Version I representations needed for learning action consequences in support of Task 4.
  Demonstrate capability, report on capability, and deliver software.
- February 2013:
  Identify representations needed for learning patterns of activity in support of Task 5.
  Deliver specification.
- September 2013:
  Complete implementation of Version I representations needed for learning patterns of activity in support of Task 5.
  Demonstrate capability, report on capability, and deliver software.
- September 2014:
  Identify representations needed for story-board learning and reuse in support of Task 6.
  Deliver specification.
- February 2015:
  Complete implementation of Version I representations needed for story-board learning in support of Task 6.
  Demonstrate capability, report on capability, and deliver software.
- September 2015:
  Complete representation specifications of all types in support of all tasks.
  Deliver specification.

Task 4: Develop commonsense rules for visual activity and means for learning such rules from questions

In our Genesis system we have well developed mechanisms handling commonsense, knee jerk, rules that record the direct consequences of physical and abstract actions. For example, if a bird flies to a tree, it makes contact with the tree; similarly, if someone kills someone else, the killed person is dead. We have also scratched the surface of what needs to be done for the Mind's Eye Program via the student-gives-a-ball-to-another-student scenario.

We need to build on this early hint-of-concept illustration to produce a robust demonstration. First, we need to develop an early list of approximately twenty activity rules to be learned. Then, we need to formulate example videos and learning-evoking questions to drive the learning process. Next, interaction needs to be live and dynamic, not via surrogate annotations as presently done. The conclusions need to be supported by multiple examples of similar events, not just a single example, with exceptional circumstances noted. Monitoring mechanisms need to be developed so that explanations are searched for when exceptions occur.

Early on, the learning will be driven by questions. Later on, the questions will be complemented by teacher and student interaction in the field of view. "Look at the man at whom your teacher is pointing."

Our approach will be to build upon the use of commonsense knowledge in the Genesis system, taking the existing software base, representations, and concepts already demonstrated to the next level. See note on evaluation at the beginning of this Work Statement.

Milestones for Task 4

- October 2011:

  Complete list of 20 activity rules and initiate work on representative examples.
  Report on analysis.
- February 2012:

  Develop evaluation metrics consistent with DARPA guidelines.
  Report on evaluation metrics.
- September 2012:

  Complete off-line learning mechanisms.
  Demonstrate learning with annotated video.
- February 2013:

  Complete on-line learning mechanisms.
  Demonstrate learning with live video.
- September 2013:

  Incorporate multiple examples in learning off line.
  Demonstrate multiple example learning with annotated video.
- February 2014:

  Incorporate multiple examples in learning on line.
  Demonstrate multiple example learning with live video.
- September 2014:

  Incorporate attention instructions into learning process.
  Demonstrate use of attention instructions.

Task 5: Develop reflective, pattern-of-activity descriptions for visual activity

In our Genesis system we have well developed mechanisms handling reflective thinking as stories are understood, so Genesis is able to recognize patterns, such as revenge, even though the defining elements may be separated by a long chain of intermediate but connected events.

We need to deploy an analog to this reflection mechanism in our Mind's Eye work. The word **bury**, for example, is a pattern of activity, rather than an atomic action: fetching a spade leads to a digging action which leads to a put action which leads to a filling action, which completes the pattern of activity; just as revenge involves a harm that leads to a desire to harm which leads to a harm.

With such a reflection mechanism in hand, the Genesis system has a ready way of directing a vision system to look for the next activity in the pattern of activity. When a vision system tells Genesis that a spade has been fetched, that observation invites Genesis to tell the vision system to focus on the spade bearer, watching particularly for digging.

Preventing an undesired consequence also is possible when something is buried or when the conditions for revenge are established, given that the early part of the pattern of activity is recognized.

Accordingly, we propose to adapt the story understanding mechanisms developed in Genesis to deal with patterns of activity, such as that exemplified by words such as *bury*, *deliver*, and *ambush*. Then, we need to develop an early list of approximately ten such patterns-of-activity.

Our approach will be to build upon the use of higher-level, reflective knowledge in the Genesis system, taking the existing software base, representations, and concepts already demonstrated to the next level. See note on evaluation at the beginning of this Work Statement.

Milestones for Task 5

- October 2011:

  Complete list of 10 patterns of activity to be learned.
  Report on list of activity patterns.
- February 2012:

  Develop evaluation metrics consistent with DARPA guidelines.
  Report on evaluation metrics.
- September 2012:

  Complete Version I system that recognizes selected activity patterns in annotated video.
  Demonstrate capability, report on capability, and deliver software.
- September 2013:
  Complete Version I system that recognizes selected activity patterns in live unannotated video.
  Demonstrate capability, report on capability, and deliver software.
- September 2014:

  Complete Version I system that learns patterns of activity from annotated video.
  Demonstrate capability, report on capability, and deliver software.

- September 2015:

  Complete Version I system that learns patterns of activity from unannotated live video. Demonstrate capability, report on capability, and deliver software.

Task 6: Develop story learning, retrieval, and reuse capability for visual activity

When you look around and describe what you see, you string situation characteristics and events together into a close analog to a story board, and collections of such visual story boards should be viewed as reusable collections of ordinary stories, ready to suggest opportunities and dangers, from story-board cues such as "Everyone is departing the square all the sudden," or "Everyone is staring at me," or "A van has never parked there before."

As such story boards accumulate, it will be valuable to be able to retrieve and reuse them as precedents, just as ordinary stories are retrieved and reused, and when common threads are seen, to extract commonsense rules and patterns of activity from collections of such story boards.

Our approach will be to build upon the use of higher-level, symbolic storyboard knowledge systems, taking the existing software base, representations, and concepts already demonstrated to the next level. See note on evaluation at the beginning of this Work Statement.

Milestones for Task 6

- October 2012:

  Complete list of 5 visual storyboards.
  Report on list of visual storyboards.
- February 2013:

  Develop evaluation metrics consistent with DARPA guidelines.
  Report on evaluation metrics.
- September 2013:

  Complete Version I system to store, index, and retrieve visual storyboards.
  Demonstrate capability, report on capability, and deliver software.
- February 2014:

  Complete Version I system to fill in missing pieces using retrieved storyboards.
  Demonstrate capability, report on capability, and deliver software.
- September 2014:

  Complete Version I system that anticipates likely evolutions using retrieved storyboards.
  Demonstrate capability, report on capability, and deliver software.
- September 2015:

  Complete Version I system that proposes interventions using retrieved situations precedents.

  Demonstrate capability, report on capability, and deliver software.

Unbudgeted Optional Work

MIT is the home of many first-tier contributors to computer vision, including William Freeman, Antonio Torralba, Berthold Horn, Tomaso Poggio, Pawan Sinha, Edward Gibson, and Shimon Ullman (part time). Today, some focus on low-level image formation and image processing; others focus on brain and cognitive science. All do relevant work and many are likely to turn their attention to questions raised by the Mind's Eye program downstream during the course of the program. Collectively, they represent a fabulous resource.

Accordingly, we list some representative downstream work opportunities, which if deemed relevant to Mind's Eye research, will enable additional contributions to visual intelligence.

Should DARPA find these opportunities to be of interest, we would be pleased to develop a budget for the additional work involved.

- Determine how to recognize actions united by name but separated by appearance. Ullman is working on how we can recognize that a cat with upturned head under a dripping fountain is drinking. Likewise, it is important to distinguish actions that are close in appearance but different in effect; a human holding a glass above his mouth is more likely to be toasting than drinking.
- Determine how we can recognize the context—beach, city, pasture, road—using statistical gisting methods developed by Torralba on large datasets.
- Determine how we can use existing statistical methods with low-resolution video, as suggested by Freeman, to recognize actions, such as those listed in the BAA.

We also anticipate that work in the Mind's Eye Program is a step toward breaking out of the coding-telling restriction, enabling progress toward the development of truly intelligent systems.

Accordingly, we list some representative downstream work opportunities, which if enabled by Mind's Eye research and eventually funded under this option, will enable DARPA contractors to deploy intelligent systems at another level.

- Determine how to support understanding of the physical world not only with visual grounding but also with grounding in motor and manipulation mechanisms.
- Integrate narrative with map and terrain understanding.
- Determine how to understand abstract use of terms (moving toward air dominance) by way of visual and motor grounding.
- Demonstrate situation retrieval based on both low-level characteristics (exhausted troops) and high-level characterizations (a enfilading movement).

2.8 Intellectual property

Identification and Assertion of Restrictions on the Government's Use, Release, or Disclosure of Technical Data or Computer Software.

The Offeror asserts for itself, or the persons identified below, that the Government's rights to use, release, or disclose the following technical data or computer software should be restricted:

The Mind of the Mind's Eye – DARPA-BAA-10-53
(Submitted by Patrick Winston)

Technical Data or Computer Software to be Furnished With Restrictions*	Basis of Assertion**	Asserted Rights Category***	Name of Person Asserting Restrictions****
Genesis System	Partially at private expense	Government purposes	Patrick Winston

* For technical data (other than computer software documentation) pertaining to items, component, or processes developed at private expense, identify both the deliverable technical data and each such items, component, or process. For computer software or computer software documentation identify the software or documentation.

** Generally, development at private expense, either exclusively or partially, is the only basis for asserting restrictions. For technical data, other than computer software documentation, development refers to development of the item, component, or process to which the data pertain. The Government's rights in computer software documentation generally may not be restricted. For computer software, development refers to the software. Indicate whether development was accomplished exclusively or partially at private expense. If development was not accomplished at private expense, or for computer software documentation, enter the specific basis for asserting restrictions.

*** Enter asserted rights category (e.g., government purpose license rights from a prior contract, rights in SBIR data generated under another contract, limited, restricted, or government purpose rights under this or a prior contract, or specially negotiated licenses).

**** Corporation, individual, or other person, as appropriate.

***** Enter "none" when all data or software will be submitted without restrictions.

Date: 7 May 2010

Printed Name and Title : Patrick Winston
Signature: Patrick Winston

(End of identification and assertion)

(Note: Offerors must complete and submit this identification and assertion as required by paragraph (d) of the Section L provisions entitled "IDENTIFICATION AND ASSERTION OF USE, RELEASE, OR DISCLOSURE RESTRICTIONS," DFARS 252.227-7017.)

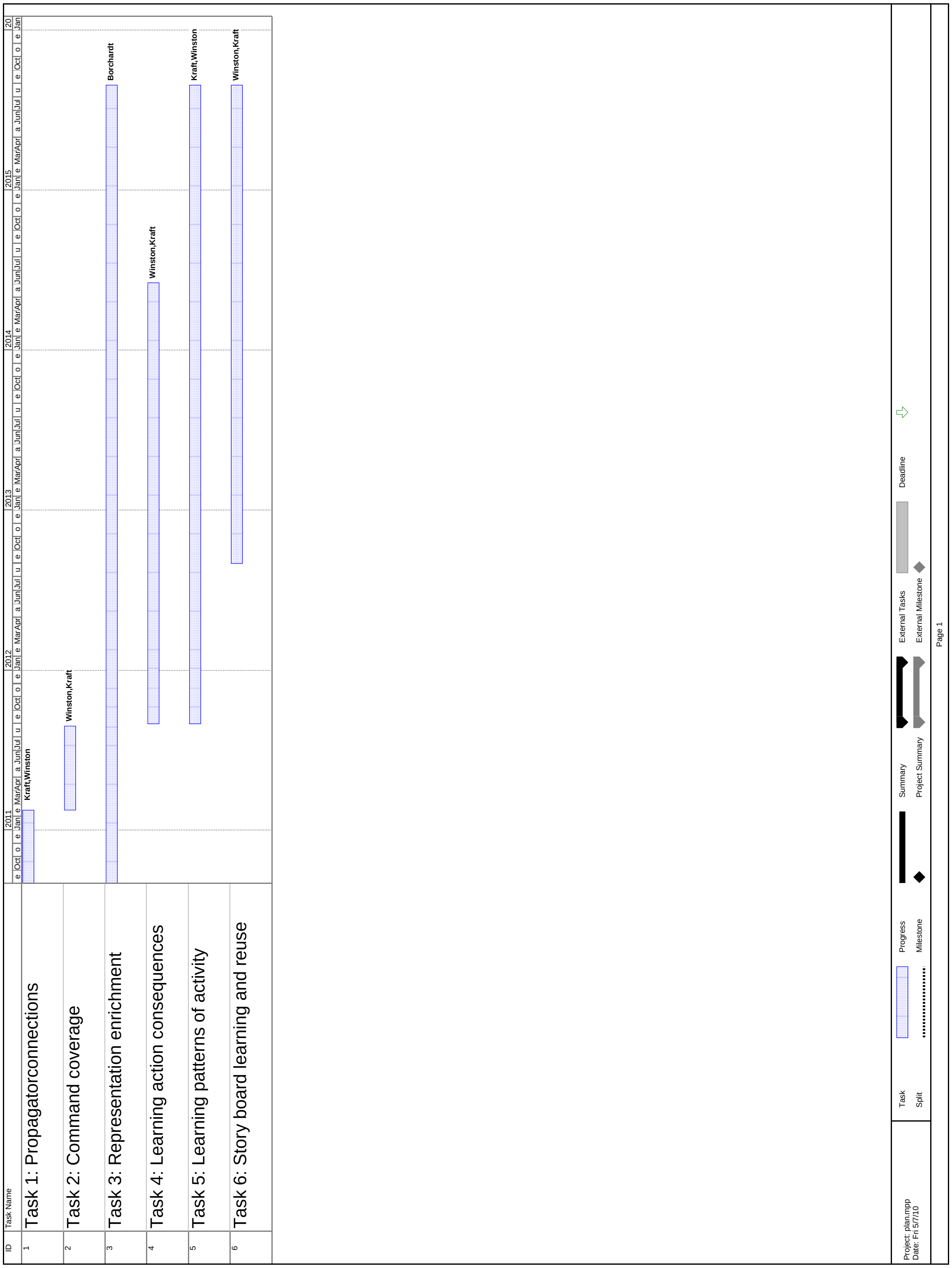
(JAN 96)

2.9 Management Plan

It is anticipated that the work proposed will be conducted in close collaboration with Co57 and/or other Mind's Eye contractors, but without a contractor–subcontractor relationship. Accordingly, no formal teaming arrangements are needed or anticipated.

Local management will be handled by the Principal Investigator via weekly staff meetings at which quarterly objectives will be set, progress reports given, and blockers resolved. It is anticipated that many unpaid students will be involved in class projects contributing to the work proposed.

2.10 Schedule and Milestones



Schedule

A project schedule is provided on the previous page. All work will proceed in parallel, so the project schedule represents focus of attention rather than stages with definite start and end dates.

Task Descriptions

Detailed task descriptions and specific milestones appear in the *Statement of Work*. Here we review responsibilities.

For Patrick Winston

Professor Winston will be the Principal Investigator, a system developer, and a student supervisor. His system development will focus on producing semantic interpretations from syntactic analysis and on interaction with visual reasoning systems, with special emphasis on Task 2, *Take command coverage to next level*, and Task 6, *Develop story learning, retrieval, and reuse capability for visual activity*. He will be responsible for progress reports, presentations, and software delivery.

For Gary Borchardt

Dr. Borchardt will focus on the centerpiece of our work, Task 3, *Enrich our representation suite*. He will develop and implement the representation extensions and additions needed by the proposed effort. He will be responsible for an evolving specification document.

For Adam Kraft

Mr. Kraft will focus on system development. He will work with Professor Winston in general and especially on Task 1, *Develop robust, always on pipes between Genesis and vision systems*. He will port our representation foundation to various programming languages used in the Mind's Eye project and provide propagator box-and-wire Internet infrastructure for those programming languages. Mr. Kraft will also work closely with Professor Winston on Task 4, *Develop commonsense rules for visual activity and means for learning such rules from questions*, and Task 5. *Develop reflective, pattern-of-activity descriptions for visual activity*.

Project Management and Interaction Plan

In the event both Co57 and this proposed work are both funded, there will be biweekly joint staff meetings at MIT or Co57.

Interaction with contractors other than Co57 will occur on an as-needed basis by email, conference call, or video conference call.

Further interaction will occur during DARPA workshops and scientific meetings.

Major software source releases to all contractors in the program will be provided via SVN or CVS semianually. Executables will be provided as produced via Webstart.

2.11 Personnel, Qualifications, Commitments

Patrick Henry Winston will be the Principal Investigator, a system developer, and a student supervisor. Professor Winston is currently in the Electrical Engineering and Computer Science Department at the Massachusetts Institute of Technology. Previously, he was a student at the East Peoria Community High School.

Professor Winston has written many books on Artificial Intelligence. He has been a champion of the idea that human intelligence is largely in our perceptual, motor, and language systems, rather than behind them.

Professor Winston has been Director of the Artificial Intelligence Laboratory (now incorporated into CSAIL), President of the Association of the Advancement of Artificial Intelligence, member of the Defence Intelligence Agency Advisory Committee, and Chair of the Naval Research Advisory Committee (NRAC). He is currently a member of NRAC and the Massport Security Advisory Committee.

Professor Winston will devote 1/3 of his research time to the proposed work while MIT is in session, which amounts to approximately one day per week. He will devote 1/2 of all his time while MIT is not in session.

Professor Winston devotes approximately 1/4 of his research time to MIT's Explorations in Cyber International Relations program (ECIR), sponsored by the Office of the Secretary of Defence, while MIT is in session, and 1/3 of all his time while MIT is not in session.

Professor Winston is currently involved in a DARPA seedling sponsored by Michael Cox on narrative understanding. The Cox seedling has no overlap with the proposed work.

Professor Winston expects to be involved in a seedling sponsored by Joseph Olive on grounding language in perception. The Olive seedling is based on prior work by Dr. Sajit Rao. Dr. Rao, now at MIT, expects to form a new enterprise, Co57, and should Co57 win a Mind's eye contract, he proposes to move to Co57 full time.

Accordingly, in the event Co57 wins a Mind's Eye contract, we will propose to have MIT subcontract the remaining work on the Olive seedling to Co57 and Professor Winston's involvement in the Olive seedling will terminate. In the event Co57 does not win a Mind's Eye contract, we anticipate that work on the Olive seedling will proceed under Dr. Rao's direction at MIT.

Because Dr. Rao is eager to continue collaboration with Professor Winston even if the proposed work at MIT is not funded, the Co57 proposal may include a consulting fee for Professor Winston. However, **In the event this proposal is accepted, Professor Winston will not accept any consulting fees from Co57** because Professor Winston's interaction with Co57 would be a natural byproduct of MIT's and Co57's participation in the program.

Dr. Gary Borchardt will focus on representation development. Dr. Borchardt completed his PhD under Professor Winston's direction and subsequently published his thesis, *Thinking Between the Lines: Computers and the Comprehension of Causal Descriptions*. The thesis introduced *transition-space*, a representation coupling visual change with device function. Transition space is a key component of the Genesis representations suite.

Dr. Borchardt will devote half his time to the proposed work. Half of Dr. Borchardt's time is currently spent on a joint project with MIT's Lincoln Laboratory aimed at understanding aerial videos of roads and streets with a view toward detecting events such as transfers and drops. This project will be concluded before or soon after the proposed work is underway.

Adam Kraft will focus on system development. His PhD thesis will focus on the role of visual intelligence in overall intelligence. His pioneering MEng thesis, done under the supervision of Professor Winston, demonstrated how programs can use examples to learn the rules need to translate from syntactic parse trees to semantic interpretations. Mr. Kraft will devote all his research-assistant time to the proposed project both while MIT is in session and while MIT is not in session.

Key Individual	Project	Pend/Curr	Year 1	Year 2	Year 3	Year 4	Year 5
Winston	Mind's Eye	Proposed	544	544	544	544	544
	Cyber	Current	388	388	388	388	n/a
	Narrative	Current	n/a	n/a	n/a	n/a	n/a
	Priming	Proposed	n/a	n/a	n/a	n/a	n/a
Borchardt	Mind's Eye	Proposed	1000	1000	1000	1000	1000
	Bluegrass	Current	n/a	n/a	n/a	n/a	n/a
Kraft*	Mind's Eye	Proposed	1000	1000	1000	1000	1000
	Teaching assistant	Current	n/a	n/a	n/a	n/a	n/a

*Adam Kraft or TBA on Kraft's graduation

Year 1	1 September 2010–31 August 2011
Year 2	1 September 2011–31 August 2012
Year 3	1 September 2012–31 August 2013
Year 4	1 September 2013–31 August 2014
Year 5	1 September 2014–31 August 2015

- Cyber

  OSD, *Explorations in Cyber International Relations*
- Narrative

  DARPA, *Defining and Demonstrating Capabilities for Experience-Based NarrativeMemory* (seedling)
- Priming

  DARPA, *Perceptual Priming for Language Learning* (seedling)
- Bluegrass:
  Lincoln Laboratory, satellite surveillance program

2.12 Organizational Conflict of Interest

None.

2.13 Human use

None.

2.14 Animal use

None.

2.15 Statement of Unique Capability

Not applicable.

2.16 Government funded Team Member Eligibility

None.

2.17 Facilities

The research will be conducted in the Computer Science and Artificial Intelligence Laboratory, MIT's largest laboratory, which provides extensive infrastructure support, including a computing cloud and a machine shop for fabricating mechanical structures. However, no resources beyond readily available computers, cameras, and high-speed network connections are contemplated for the proposed research.