# Clashes in the Infosphere, General Intelligence, and Metacognition: Final project report

Computer Science Technical Report No. CS-TR-5017

UMIACS Technical Report No. UMIACS-TR-2012-12

**Don Perlis[1] and Michael T. Cox[2]**

perlis@cs.umd.edu

[1]Department of Computer Science, University of Maryland

[2]University of Maryland Institute for Advanced Computer Studies (UMIACS)

College Park, MD 20742

**Abstract**

Humans confront the unexpected every day, deal with it, and often learn from it. AI agents, on the other hand, are typically brittle—they tend to break down as soon as something happens for which their creators did not explicitly anticipate. The central focus of our research project is this *problem of brittleness* which may also be the single most important problem facing AI research. Our approach to brittleness is to model a common method that humans use to deal with the unexpected, namely to *note* occurrences of the unexpected (i.e., anomalies), to *assess* any problem signaled by the anomaly, and then to *guide* a response or solution that resolves it. The result is the Note-Assess-Guide procedure of what we call the Metacognitive Loop or MCL. To do this, we have implemented MCL-based systems that enable agents to help themselves; they must establish expectations and monitor them, note failed expectations, assess their causes, and then choose appropriate responses. Activities for this project have developed and refined a human-dialog agent and a robot navigation system to test the generality of this approach.

## Activities

The research has focused on testing the hypothesis that a properly designed metacognitive loop can be both domain- and anomaly-general, by implementing it in a (partly physical, partly simulated) world containing many of the most challenging and important facets of the general infosphere problem, which involves multiple autonomous agents, which can gather and transmit data as well as request, receive and process data. We have, in fact, implemented the Metacognitive Loop in several distinct types of systems: a reinforcement learner, a human-computer dialogue agent, a tank game, a robot navigation system, and a commonsense (nonmonotonic) reasoner. In each case, the performance of the system was enhanced by MCL mechanisms. Activities for the current project have developed and refined a human-dialog agent and a robot navigation system. The former is called ALFRED and the latter the simulated Mars rover.

ALFRED is a universal interfacing agent that accepts a set of English sentences and translates them into commands appropriate for different domains. The original ALFRED had no parsing capacity (parsing was conducted by an outside, off-the-shelf system), and its knowledge of language in general was extremely limited. In this project we developed ALFRED 2.0, the next generation of this interfacing agent. ALFRED 2.0, like its predecessor, is written in Alma, a Prolog-like, declarative language (Alma itself is written in Prolog). What makes Alma unique is its step-wise temporal reasoning capability. Alma keeps track of time by counting off steps (or time units). Each step in a logical derivation corresponds to an Alma step.

For instance, at step 1, Alma can accept that P, P -> Q, and Q -> R is true. Then each step in the derivation of R takes a time step. In a language like Prolog, R would be concluded immediately instead. The step-by-step reasoning was designed to accommodate contradictions in an agent's knowledge base. Thus, ALFRED 2.0 is a semantic ontology, i.e. a characterization of knowledge, written in the Alma formalism. Alma is the reasoning engine that, step-by-step, expands/changes ALFRED's knowledge base. This knowledge consists of linguistic information (mostly) but was designed so that general knowledge could be added easily. ALFRED 2.0 was created to provide a natural language, bidirectional interface to an open set of domains. It acts as a translator, taking typed English utterances as input (although it can also take voice input from a speech-to-text software package called Dragon), and outputs the domain appropriate command string.

We have also developed a substrate for simulating discrete worlds in which metacognitive agents can act. Upon this substrate, we have implemented a basic Mars domain in which simulated Mars Rovers can be deployed to conduct basic scientific missions. The simulated Rovers possess planners for bottom-up navigation and for high-level mission activities, and they have basic monitor and control capabilities that are integrated with MCL. Coupled with control over perturbations to the domain, the Mars Rover provides a simple, flexible, and extensible platform for developing the monitor and control portions of the MCL ontologies and a test bed for demonstrating the general efficacy of MCL. Further, the Mars domain provides basic TCP/IP command-dispatching agents and goal monitoring agents to provide a common interface for introducing goals and perturbations to the simulated environment and scoring the progress of Mars Rovers, respectively. We also have configured the same MCL software to connect to

physical mobile robots (at the Naval Research Lab, at Franklin & Marshall College, and at the University of Maryland, Baltimore County).

Finally, we have an overarching hypothesis: that the identical core MCL code will serve to guide multiple applications, and in others still to be designed. As such MCL is a monitoring-and-control module that can be attached to a given "host" system, thereby yielding an improved system. The team has designed and built a prototype version of generalized MCL and attached it to various host systems including search-and-rescue robots and Mars-style rovers (both real and simulated), a simulated air traffic control system, and a natural-language-based human-computer interaction system.

Testing our overarching hypothesis requires producing ontologies of indications, failures, and repairs for these domains and extracting common core structure. We did this for human-computer dialog systems with the goal of allowing the computer to automatically determine when the human is dissatisfied with the dialog and to effect a repair. The ontologies were constructed by reviewing the literatures on human and human-computer dialog, and by running experiments with human subjects interacting with AIML chatbots. The output of the process was a complete set of ontologies, with linkages both within and between ontologies suitable for ingest by the MCL implementation.

We also developed simulated search and rescue robots that can avoid obstacles, locate items of interest and gather these items. An increase in number of collisions (encounters with danger) causes an expectation violation that a metacognitive component notes. This expectation violation causes the metacognitive component to adjust different parameters like the safe distance from obstacle, and the speed and time spend on path planning. The ontology of indications, failures and responses that the on-board, metacognitive component of the robots use is compatible with MCL.

We have built specialized controllers for two Corobots, integrated the controllers with the central communication system ("RonCon"), and designed and deployed appropriate test courses for the robots. We also designed and implemented a new algorithm for color classification that shows significant robustness across lighting conditions. This is currently being prepared for publication.

At UMBC we developed the core ontology-based MCL implementation, the monitor and control ("MonCon") infrastructure that connects distributed host systems and humans to explore the behavior of MCL in multi-agent environments, and the analogous infrastructure (RonCon) suitable for robotic agents. We also implemented the Mars Rover domain and integrated it with MonCon. Finally, the work with AIML chatbots, including human subjects experiments, was done at UMBC, as was most of the development of the dialog ontologies for MCL.

Underlying the MCL approach is a time-sensitive, contradiction-tolerant logical reasoning engine called the active logic machine (ALMA). We have formalized a semantics for a general version of the underlying logical formalism, active logic. Central to active logic are special rules controlling the inheritance of beliefs in general (and of beliefs about the current time in particular), very tight controls on what can be derived from direct contradictions (P&¬P), and mechanisms allowing an agent to represent and reason about its own beliefs and past reasoning. Furthermore, inspired by the notion that until an agent notices that a set of beliefs is

contradictory, that set seems consistent (and the agent therefore reasons with it as if it were consistent), we introduce an "apperception function" that represents an agent's limited awareness of its own beliefs, and serves to modify inconsistent belief sets so as to yield consistent sets. Using these ideas, we introduced a new definition of logical consequence in the context of active logic, as well as a new definition of soundness such that, when reasoning with consistent premises, all classically sound rules remain sound in our new sense. However, not everything that is classically sound remains sound in our sense, for by classical definitions, all rules with contradictory premises are vacuously sound, whereas in active logic not everything follows from a contradiction.

The process of rationally revising beliefs in the light of new information is a topic of great importance and long-standing interest in artificial intelligence. Moreover, significant progress has been made in understanding the philosophical, logical, and computational foundations of belief revision. However, very little research has been reported with respect to the revision of other mental states, most notably propositional attitudes such as desires and intentions. In this project, we presented a first attempt to formulate a general framework for understanding the revision of mental states. We developed an abstract belief-desire-intention model of agents, and introduce a notion of rationality for this model. We then presented a series of formal postulates characterizing the processes of adding beliefs, desires, and intentions, updating costs and values, and removing beliefs, desires, and intentions. We also investigated the computational complexity of several problems involving this abstract model.

We have also made advances in MCL-related natural language understanding. Language not only refers to objects and events in the world, but it also can refer to language constituents themselves. This metalinguistic phenomena is captured in the *use-mention distinction*. In particular, we are able to recognize mentioned language: that is, tokens (e.g., words, phrases, sentences, letters, symbols, sounds) produced to draw attention to linguistic properties that they possess. Evidence suggests that humans frequently employ the use-mention distinction, and we would be severely handicapped without it; mentioned language frequently occurs for the introduction of new words, attribution of statements, explanation of meaning, and assignment of names. Moreover, just as we benefit from mutual recognition of the use-mention distinction, the potential exists for us to benefit from language technologies that recognize it as well. With a better understanding of the use-mention distinction, applications can be built to extract valuable information from mentioned language, leading to better language learning materials, precise dictionary building tools, and highly adaptive computer dialogue systems.

Three specific contributions were made. The first is a framework for identifying and analyzing instances of mentioned language, in an effort to reconcile elements of previous theoretical work for practical use. Definitions for mentioned language, metalanguage, and quotation have been formulated, and a procedural rubric has been constructed for labeling instances of mentioned language. The second is a sequence of three labeled corpora of mentioned language, containing delineated instances of the phenomenon. The corpora illustrate the variety of mentioned language, and they enable analysis of how the phenomenon relates to sentence structure. Using these corpora, inter-annotator agreement studies have quantified the concurrence of human readers in labeling the phenomenon. The third contribution is a method for identifying common forms of mentioned language in text, using patterns in metalanguage and sentence structure. Although the full breadth of the phenomenon is likely to elude computational tools for the

foreseeable future, some specific, common rules for detecting and delineating mentioned language have been shown to perform well.

Given these results, we are beginning two new research thrusts to extend the current research reported here. First we have started to examine the psychological and cognitive neuroscience correlates to human metacognitive processes similar to the MCL mechanism. Initial studies by Anderson at Franklin & Marshall have examined neural reuse and functional connectivity in the brain and how all of cognition, even metacognition, is grounded in concrete experience. Second we have begun a long-term project to apply our results to the design of a large cognitive architecture that focuses upon the interaction of cognition and metacognition. The system is called MIDCA for the Metacognitive Integrated Dual-Cycle Architecture. These two new projects show much early promise and would not have come about without the foundational research this AFOSR funding provided.

## Contributions

This work modifies and enhances decades-long efforts to design logic-based commonsense-reasoning intelligent systems. Chiefly, we incorporated real-time, natural-language, and inconsistency-tolerant aspects into the system capabilities, by means of a re-conceptualization of common sense. Namely, we investigated autonomous error-correction (note, assess, and respond to anomalies) as a key element heretofore missing in this area. Systems so-enhanced appear to be far more robust and adaptive in the face of unanticipated circumstances.

## Training and Development

Various students and postdocs were fully enmeshed in this work. Both the development and the testing of the MCL system has allowed us to engage students at many different levels of the research, from theoretical work, to programming, to physically setting up robotic test courses. In addition, many students participated in high-level conceptual work as the project evolved to later stages. Three PhD students finished their research with support from this grant and have successfully graduated. Two postdoctoral fellows received valuable guidance under this grant and contributed significantly to the research program. The high-school students involved in the project finished their parts of the project successfully, in the process learning a great deal about AI, some of it quite different from what they had imagined.

## Outreach

During early 2011, we established a new seminar series at the University of Maryland College Park campus. The *Maryland Metacognition Seminar* (www.cs.umd.edu/active/MetaCogSeminar) reaches out to institutions in the greater DC metropolitan area to attract faculty and students that engage or are interested in research related to important issues in human and machine metacognition and metareasoning. Don Perlis gave the inaugural presentation on March 4, 2011, and thirteen additional research presentations followed up to the current time. The ongoing series provides an interdisciplinary forum for both preliminary and mature research results and normally occurs on a monthly basis.

The results of our research and all of the publications that have originated with this grant are available on the web for public release. See www.cs.umd.edu/projects/active.

## Acknowledgements

This research is funded by the Air Force Office of Scientific Research (AFOSR/RSL) under Grant No. FA9550-09-1-0144. The period of performance is 01Mar2009 through 30Nov2012. The Principal Investigator is Don Perlis.

## Project References

### Theses

Kelley, T. (2012). *The functional connectivity of the brain under metacognition*. Senior Honors Thesis, Department of Psychology, Franklin & Marshall College, Lancaster, Pennsylvania.

Haidarian, H. (2011). *On the foundations of data interoperability and semantic search on the web*. PhD dissertation. University of Maryland, College Park, MD.

Wilson. S. (2011). *A computational theory of the use-mention distinction in natural language*. PhD dissertation. University of Maryland, College Park, MD.

Wright, D. (2011). *Finding a temporal comparison function for the metacognitive loop*. PhD dissertation. University of Maryland, Baltimore County, Baltimore, MD.

### Journals and Magazines

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences 33*(4), 245-66.

Anderson, M. L. (2008). Circuit sharing and the implementation of intelligent systems. *Connection Science 20*(4), 239-251.

Anderson, M. L., Fults, S., Josyula, D. P., Oates, T., Perlis, D., Schmill, M. D., Wilson, S., & Wright, D. (2008). A self-help guide for autonomous systems. *AI Magazine 29*(2): 67-76.

Anderson, M., Gomaa, W., Grant, J., & Perlis, D. (2008). Active logic semantics for a single agent in a static world. *Artificial Intelligence 172*, 1045-1063.

Anderson, M. L., & Perlis, D. (2009). What puts the "meta" in metacognition? *Behavioral and Brain Sciences 32*(2), 138-139.

Cox, M. T., Oates, T., Paisner, M., & Perlis, D. (in press). *Noting anomalies in streams of symbolic predicates using A-distance*. To appear in *Advances in Cognitive Systems*.

Cox, M. T., & Perlis, D. (2011, November). Self-adjusting autonomous systems. *Awareness Magazine*. DOI: 10.2417/3201111.003951.

Grant, J., Kraus, S., Perlis, D., Wooldridge, M. (2010). Postulates for revising BDI structures. *Synthese 175*, 39–62.

Perlis, D. (2010). To BICA and beyond: How biology and anomalies together contribute to flexible cognition. *International Journal of Machine Consciousness 2*(2), 1-11.

Wilson, S. (2011). In search of the use-mention distinction and its impact on language processing tasks. *International Journal of Computational Linguistics and Applications 2*(1-2), 139-154.

## Book Chapters

Anderson, M. L. (2008). On the grounds of x-grounded cognition. In P. Calvo & T. Gomila (Eds.), *The Elsevier Handbook of Cognitive Science: An Embodied Approach* (pp. 423-435). Amsterdam: Elsevier Science.

Anderson, M. L., Gomaa, W., Grant, J. & Perlis, D. (in press). An approach to human-level commonsense reasoning. In F. Berto, E. Mares, F. Paoli & K. Tanaka (Eds.), *Paraconsistent Logic*.

Cox, M. T. (2011). Metareasoning, monitoring, and self-explanation. In M. T. Cox & A. Raja (Eds.) *Metareasoning: Thinking about thinking* (pp. 131-149). Cambridge, MA: MIT Press.

Cox, M. T., & Raja, A. (2011). Metareasoning: An introduction. In M. T. Cox & A. Raja (Eds.) *Metareasoning: Thinking about thinking* (pp. 3-14). Cambridge, MA: MIT Press.

Fults, S. (2011). Vagueness and scales. In P. Egre and N. Klinedinst (Eds.), *Vagueness and language use* (pp. 25-49). Basingstoke, UK: Palgrave Macmillan.

Gordon, A. S., Hobbs, J. R., & Cox, M. T. (2011). Anthropomorphic self-models for metareasoning agents. In M. T. Cox & A. Raja (Eds.) *Metareasoning: Thinking about thinking* (pp. 295-305). Cambridge, MA: MIT Press.

Perlis, D. (2011). There's no ``Me" in ``Meta" - or is there? In M. T. Cox & A. Raja (Eds.) *Metareasoning: Thinking about thinking* (pp. 15-26). Cambridge, MA: MIT Press.

Schmill, M. D., Anderson, M. L., Fults, S., Josyula, D., Oates, T., Perlis, D., Haidarian, H., Wilson, S., & Wright, D. (2011). The metacognitive loop and reasoning about anomalies. In M. T. Cox & A. Raja (Eds.), *Metareasoning: Thinking about thinking* (pp. 183-198). Cambridge, MA: MIT Press.

## Proceedings

Anderson, M. L., & Aktipis, C. A. (2010). The origins of collective overvaluation: Irrational exuberance emerges from simple, honest and rational individual behavior. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.

Anderson, M. L., & Oates, T. (2010). A critique of multi-voxel pattern analysis. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (p. 1511-1516). Austin, TX: Cognitive Science Society.

Bhargava, P., Cox, M. T., Oates, T., Oh, U., Paisner, M., Perlis, D., & Shamwell, J. (in press). The robot baby and massive metacognition: Future vision. To appear in *Proceedings of the IEEE Conference on Development and Learning - Epigenetic Robotics 2012 (ICDL/EpiRob)*.

Cox, M. T., Oates, T., Paisner, M., & Perlis, D. (2012). *Detecting change in diverse symbolic worlds*. Submitted to AAMAS-13 Conference.

Cox, M. T., Oates, T., & Perlis, D. (2011). Toward an integrated metacognitive architecture. In P. Langley (Ed.), *Advances in Cognitive Systems, papers from the 2011 AAAI Symposium* (pp. 74-81). Technical Report FS-11-01. Menlo Park, CA: AAAI Press.

Gold, K., Havasi, C., Anderson, M. L., & Arnold, K. (2011). Comparing matrix decomposition methods for meta-analysis and reconstruction of cognitive neuroscience results. In *Proceedings of the 24th Annual Conference of the Florida Artificial Intelligence Research Society (FLAIRS-24)*.

Haidarian, H. (2010a). Foundations of Data Interoperability on the Web: A Web Science Perspective. In *Proceedings of the 6th International Conference on Semantic Systems (I-Semantics'10)*, Graz, Austria, September 1-3.

Haidarian, H. (2010b). Semantic Search in Linked Data: Opportunities and Challenges. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI'10)*, Atlanta, Georgia, USA, July 11-15.

Haidarian, H., Dinalankara, W., Fults, S., Wilson, S., Perlis, D., Schmill, M., Oates, T., Josyula, D. & Anderson, M. (2010). The metacognitive loop: An architecture for building robust intelligent systems. In *Proceedings of the AAAI Fall Symposium on Commonsense Knowledge*, Arlington, VA, USA, November 11-13.

Haidarian, H., & Perlis, D. (2008). Finding ontological correspondences for a domain-independent natural language dialog agent. In *Proceedings of the 20th AAAI Innovative Applications of Artificial Intelligence Conference (AAAI/IAAI'08)*, Chicago, USA, July 13-17. Menlo Park, CA: AAAI Press.

Josyula, D. P., Donahue, B., McCaslin, M., Snowden, M., Anderson, M., Schmill, M., Oates, T. & Perlis, D. (2010). Metacognition for detecting and resolving conflicts in operational policies. In *Proceedings of the Workshop on Metacognition for Robust Social Systems at the 24th AAAI Conference on Artificial Intelligence*, Atlanta, Georgia, USA, July 11-15.

Josyula, D. P., Hughes, F. C., Vadali, H., & Donahue, B. J. (2009). Modeling emotions for choosing between deliberation and action. In *Proceedings of the IEEE World Congress on Nature and Biologically Inspired Computing (NABIC'09)*.

Josyula, D. P., Hughes, F. C. Vadali, H., Donahue, B. J., Molla, F., Snowden, M., Miles, J., Kamara, A. & Maduka, C. (2009). Metacognition for self-regulated learning in a dynamic environment. In *Proceedings of the 2009 SASO Workshop on Metareasoning in Self-Adaptive System, at the Third IEEE International Conference on Self-Adaptive and Self-Organizing Systems (SASO'09)*.

Josyula, D. P., Vadali, H., Donahue, B. J., & Hughes, F. C. (2009). Modeling metacognition for learning in artificial systems. In *Proceedings of the International Symposium on Innovations in Computing (INC'09)*.

Shamwell, J., Oates, T., Bhargava, P., Cox, M. T., Oh, U., Paisner, M., & Perlis, D. (in press). The robot baby and massive metacognition: Early steps via growing neural gas. To appear in *Proceedings of the IEEE Conference on Development and Learning - Epigenetic Robotics 2012 (ICDL/EpiRob)*.

Subramanian, A., & Oates, T. (2009). Ontologies for metacognitive monitoring and repair of dialog. In *Proceedings of the 4th Language and Technology Conference*, Poznan, Poland, November 6-8, 2009.

Subramanian, A., Oates, T., & Fults, S. (2010). Ontologies for monitoring and repairing human-computer dialogs. In *Papers from the 6th IEEE International Conference on Natural Language Processing and Knowledge Engineering*, Beijing, China, Aug 21- 23.

Wilson, S. (2012). The creation of a corpus of English metalanguage. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (pp. 638–646). Jeju, Republic of Korea, 8-14 July. Association for Computational Linguistics.

Wilson, S. (2010). Distinguishing use and mention in natural language. In Proceedings of the NAACL HLT Student Research Workshop (pp. 29-33). Los Angeles, CA: Association for Computational Linguistics.

## Talks

Josyula, D. (2010). *Cracking the brittleness problem-path to robust intelligence*. Invited talk at Mar Baselios College of Engineering and Technology, Thiruvananthapuram, India, August 19.

Perlis, D. (2008). *To BICA and beyond: How biology and anomalies together contribute to flexible cognition.* Keynote address, AAAI Fall Symposium on Biologically Inspired Cognitive Architectures, Washington DC.

Perlis, D. (2008). *There's no ``Me'' in ``Meta'' - or is there?* D. Perlis. Keynote address, AAAI Workshop on Metareasoning: Thinking about thinking Workshop, AAAI Annual Conference, Chicago.