2010

A Conceptual and Computational Model of Moral Decision Making in Human and Artificial Agents

W. Wallach

S. Franklin

C. Allen

Recommended Citation
Wallach, W., Franklin, S., & Allen, C. (2010). A Conceptual and Computational Model of Moral Decision Making in Human and Artificial Agents. 10.1111/j.1756-8765.2010.01095.x

A Conceptual and Computational Model of Moral Decision Making in Human and Artificial Agents

By Wendell Wallach, Stan Franklin, and Colin Allen

Abstract: Recently there has been a resurgence of interest in general, comprehensive models of human cognition. Such models aim to explain higher order cognitive faculties, such as deliberation and planning. Given a computational representation, the validity of these models can be tested in computer simulations such as software agents or embodied robots. The push to implement computational models of this kind has created the field of Artificial General Intelligence, or AGI.

Moral decision making is arguably one of the most challenging tasks for computational approaches to higher order cognition. The need for increasingly autonomous artificial agents to factor moral considerations into their choices and actions has given rise to another new field of inquiry variously known as Machine Morality, Machine Ethics, Roboethics or Friendly AI. In this paper we discuss how LIDA, an AGI model of human cognition, can be adapted to model both affective and rational features of moral decision making. Using the LIDA model we will demonstrate how moral decisions can be made in many domains using the same mechanisms that enable general decision making.

Comprehensive models of human cognition typically aim for compatibility with recent research in the cognitive and neural sciences. Global Workspace Theory (GWT), proposed by the neuropsychologist Bernard Baars (1988), is a highly regarded model of human cognition that is currently being computationally instantiated in several software implementations. LIDA (Franklin et al. 2005) is one such computational implementation. LIDA is both a set of computational tools and an underlying model of human cognition, which provides mechanisms that are capable of explaining how an agent's selection of its next action arises from bottom-up collection of sensory data and top-down processes for making sense of its current situation. We will describe how the LIDA model helps integrate emotions into the human decision making process, and elucidate a process whereby an agent can work through an ethical problem to reach a solution that takes account of ethically relevant factors.

Keywords: moral decision making, artificial general intelligence, artificial intelligence, Global Workspace Theory, machine morality, machine ethics

Introduction

AGI and Moral Machines

Human-level intelligence entails the capacity to handle a broad array of challenges including logical reasoning, understanding the semantic content of language, learning, navigating around the obstacles in a room, discerning the intent of other agents, and planning and decision making in situations where information is incomplete. The prospect of building "thinking machines" with the general intelligence to tackle such an array of tasks inspired the early founders of the field of Artificial Intelligence. However, they soon discovered that tasks such as reasoning about physical objects or processing natural language, where they expected to make rapid progress, posed daunting technological problems. Thus the developers of AI systems have been forced to focus on the design of systems with the ability to intelligently manage specific tasks within relatively narrow domains, such as playing chess or buying and selling currencies on international markets. Despite the fact that many tasks such as visual processing, speech processing, and semantic understanding present thresholds that have yet to be crossed by technology, there has been in recent years a transition back to the development of systems with more general intelligence. Such systems are broadly referred to as having artificial general intelligence (AGI) (Wang *et* al. 2008, Goertzel and Franklin 2008).

The possibility of building AI systems with moral decision making faculties has stepped beyond the stories of science fiction writers such as Isaac Asimov and is being seriously considered by philosophers and engineers, (Gips 1991, Clark 1994, Allen et al. 2000). A new field of enquiry called Machine Ethics (Anderson and Anderson 2006), Machine Morality (Wallach et al. 2008), Artificial Morality (Danielson 1992), or Computational Ethics (Allen 2002) is emerging.

This interest in building computer systems capable of making moral decisions (“moral machines”) has been spurred by the need to insure that increasingly autonomous computer systems and robots do not cause harm to humans and other agents worthy of moral consideration (Wallach and Allen 2008). Though the goals of this new research endeavor are more practical than theoretical, an interest in testing whether consequentialist, deontological, and virtue based theories of ethics can be implemented computationally has also attracted philosophers and social scientists to this new field. Most of the research to date is directed at either the safety of computers that function within very limited domains or at systems that serve as advisors to human decision makers.

AGI and Machine Morality have emerged as distinct fields of inquiry. The

intersection between their agendas has been minimal, and primarily focused on Friendly AI (Yudkovsky 2001), the concern that future super-intelligent machines be friendly to human. But let us be clear at the outset. No AGI systems have been completed. Nor do any computers systems exist that are capable of making sophisticated moral decisions. However, some computer scientists believe such systems can be build relatively soon. Ben Goertzel estimates that, with adequate funding, scientists could complete an AGI within ten years (personal communication). Certainly sophisticated moral machines will require at least a minimal AGI architecture.

So, if demonstrated success in either of these pursuits is so far in the future, what do we expect to achieve in this paper. Our goals are twofold:

1. To outline a comprehensive approach to moral decision making. Philosophers and cognitive scientists have stressed the importance of particular cognitive mechanisms, e.g., reasoning, moral sentiments, heuristics, intuitions, or a moral grammar in the making of moral decisions. But there has been very little work on thinking comprehensively about the broad array of cognitive faculties necessary for moral decision making. In analyzing how a moral machine might be built from the ground up, it becomes apparent that many cognitive mechanisms must be enlisted to produce judgments sensitive to the considerations humans accommodate when they respond to morally charged situations (Wallach and Allen 2008).

2. To demonstrate that many moral decisions can be made using the same cognitive mechanisms that are used for general decision making. In other words, moral cognition is supported by domain general cognitive processes. Certainly some kinds of moral decisions may require additional mechanisms, or may require that the kinds of mechanisms described in this paper be modified to handle features peculiar to moral considerations. Elucidation of such mechanisms and their probable design is beyond the scope of this paper.

In proposing a comprehensive model for moral decision making, we are fully aware that other scholars will criticize this model as being inadequate. For example, neuroscientists might argue that a modular system such as LIDA does not capture the full complexity of the human neural architecture. Moral philosophers might contend that the agent we will describe is not really engaged in moral reflection because it lacks Kantian 'autonomy' or 'will'. The computer scientist Drew Mcdermott (forthcoming) asserts that appreciating the tension between self-interest and the needs of others is essential for moral decisions and will be extremely difficult to build into computational agents. There are many

criticisms that can be made of AGI models, and many arguments as to why computational agents are not capable of 'true' moral reflection.

Nevertheless, we feel it is important to recognize that moral judgment and behavior is not the product of one or two dedicated mechanisms. Nor do we feel it is helpful to merely underscore the complexity of moral decision making. Therefore, we offer this model in hopes of stimulating a deeper appreciation of the many cognitive mechanisms that contribute to the making of moral decisions, and to provide some insight into how these mechanisms might work together.

Computation Models of Human Cognitive Faculties

A central fascination with AI research has been the opportunity it offers to test computational theories of human cognitive faculties. AGI does not require that the computational system emulate the mechanisms of human cognition in order to achieve a comparable level of performance. However, human cognition is the only model we currently have for general intelligence or moral decision making (although some animals demonstrate higher order cognitive faculties and pro-social behavior). The cognitive and brain sciences are bringing forth a wealth of empirical data about the design of the human nervous system, and about human mental faculties. This research suggests a host of new theories for specific cognitive processes that can, at least in principle, be tested computationally.

Despite significant gaps in scientific understanding it is feasible to design systems that try to emulate the current best understanding of human faculties, even if those systems don't perform exactly as the brain functions. Computational models of human cognition are built by computer scientists who wish to instantiate human-level faculties in AI, and by cognitive scientists and neuroscientists formulating testable hypotheses compatible with empirical data from studies of the nervous system, and mental and behavioral activity.

Both as a set of computational tools and an underlying model of human cognition, LIDA is one attempt to computationally instantiate Baars' Global Workspace Theory (GWT). Such a computational instantiation of GWT, which attempts to accommodate the psychological and neuroscientific evidence, will be particularly helpful in thinking through an array of challenges with a high degree of specificity. In this paper, we will explore how the LIDA model of GWT can be expected to implement a higher order cognitive task, specifically the kind of decision making involved in the resolution of a moral dilemma.

Given that computational approaches to moral decision making, GWT, and the LIDA model are subjects that may not be familiar to all readers, the initial

sections of this paper provide brief overviews of these topics. The next section of the paper introduces several approaches to computerizing ethics, GWT, and LIDA. The following section provides a description of the LIDA model, and various theories and research that support this approach to human cognition. A discussion of the manner in which the LIDA model might be used to make moral decisions and some concluding comments follow.

Machine Morality, GWT, and LIDA

Ethical Decision Making and AI

Ethical decisions are among the more complex that agents face. Ethical decision-making can be understood as action selection under conditions where constraints, principles, values, and social norms play a central role in determining which behavioral attitudes and responses are acceptable. Many ethical decisions require having to select an action when information is unclear, incomplete, confusing, and even false, where the possible results of an action cannot be predicted with any significant degree of certainty, and where conflicting values can inform the decision making process.

Commonly, ethics is understood as focusing on the most intractable of social and personal challenges. Debate often centers on how to prioritize duties, rules, or principles when they conflict. But ethical factors influence a much broader array of decisions than those we deliberate upon as individuals or as a community. Values and ideals are instantiated in habits, normative behavior, feelings, and attitudes. Ethical behavior includes not only the choices we deliberate upon, but also the rapid choices that substantiate values – choices that might be modeled in LIDA as single-cycle, consciously mediated responses to challenges. Given this broad definition of ethical decisions, values play an implicit role, and sometimes an explicit role, in the selection of a broad array of actions.

Following Sloman (1999), we note that moral behavior can be reflexive, or the result of deliberation, and at least for humans, also includes metacognition[1] when criteria used to make ethical decisions are periodically reevaluated. Successful responses to challenges reinforce the selected behaviors, while unsuccessful outcomes have an inhibitory influence, and may initiate a reinspection of one's actions and behavior selection. Thus, a computational model of moral decision making will need to describe a method for implementing reflexive value laden responses, while also explaining how these responses can be reinforced, or inhibited through learning, top-down deliberative reasoning, and metacognition.

[1] Sloman speaks of meta-management rather than metacognition. We prefer the more common psychological term.

It is helpful, although somewhat simplistic, to think of implementing moral decision making faculties in AI systems in terms of two approaches: top-down and bottom up (Allen et al. 2000, Wallach et al. 2008, Wallach and Allen, 2008). A top-down approach entails the implementation of rules or a moral theory, such as the Ten Commandments, Kant's categorical imperative, Mill's utilitarianism, or even Asimov's laws. Generally, top-down theories are deliberative and even metacognitive, although individual duties may be implemented reactively. A top-down approach takes an antecedently specified ethical theory and analyzes its computational requirements to guide the design of algorithms and subsystems capable of implementing the theory.

A number of scholars have considered the challenges entailed in computational implementation of individual top down theories of ethics, including Asimov's laws (Clark 1994), Kant's categorical imperative (Allen et al. 2000, Stahl 2002, Powers 2005), Ross's prima facie duties (Anderson et al. 2005, 2006), deontic logic (Bringsford et al. 2006), utilitarianism (Allen et al. 2000, Grau 2006) and virtues (Demoss 1998). Implementing each of these theories poses specific difficulties for designers and programmers. Each is susceptible to some version of the frame problem -- computational load due to the need for knowledge of human psychology, knowledge of the affects of actions in the world, and the difficulty in estimating the sufficiency of initial information.

Bottom-up approaches, if they use a *prior* theory at all, do so only as a way of specifying the task for the system, but not as a way of specifying an implementation method or control structure. A bottom-up approach aims at goals or standards that may or may not be specified in explicit theoretical terms. Evolution, development, and learning provide models for designing systems from the bottom up. Alife (artificial life) experiments within computer environments, evolutionary and behavior based robots, and genetic algorithms all provide mechanisms for building sophisticated computational agents from the bottom up. Bottom-up strategies influenced by theories of development are largely dependent on the learning capabilities of artificial agents. However, the bottom-up development of moral agents is limited given present day technologies, but breakthroughs in computer learning or Alife, for example, might well enhance the usefulness of these platforms for developing artificial moral agents (Wallach and Allen 2008).

Furthermore, even agents who adhere to a deontological ethic or are utilitarians may require emotional intelligence as well as other “supra-rational” faculties (Wallach and Allen 2008). A sense of self, a theory of mind (ToM), an appreciation for the semantic content of information, and functional (if not phenomenal) consciousness (Franklin, 2003) are probably also prerequisites for full moral agency. A complete model of moral cognition will need to explain how such faculties are represented in the system.

Work has begun on the development of artificial mechanisms that complement a

system's rational faculties, such as affective skills (Picard 1997), sociability (Breazeal 2002), embodied cognition (Brooks 2002; Glenberg 1997), theory of mind (Scasselatti 2001) and consciousness (Holland 2003), but these projects are not specifically directed at designing systems with moral decision making faculties. Eventually there will be a need for hybrid systems that maintain the dynamic and flexible morality of bottom-up systems, which accommodate diverse inputs, while subjecting the evaluation of choices and actions to top-down principles that represent ideals we strive to meet. Depending upon the environments in which these artificial moral agents (AMAs) operate, they will also require some additional supra-rational faculties. Such systems must also specify just how the bottom-up and top-down processes interact.

To date, the experimental systems that implement some sensitivity to moral considerations (McLaren 2006, Anderson 2006, Guarini 2006) are rudimentary, and cannot accommodate the complexity of human decision making. Scaling any approach to handle more and more difficult challenges will, in all likelihood, require additional mechanisms.

Global Workspace Theory

Global workspace theory (GWT) (Baars 1988) was originally conceived as a neuropsychological model of consciousness, but has come to be widely recognized as a high-level theory of human cognitive processing, which is well supported by empirical studies (Baars 2002). GWT views the nervous system as a distributed parallel system with many different specialized processes. Some coalitions of these processes enable the agent to make sense of the sensory data coming from the current environmental situation. Other coalitions incorporating the results of the processing of sensory data compete for attention. The winner occupies what Baars calls a global workspace, whose contents are broadcast to all other processes. These contents of the global workspace are presumed to be conscious, at least from a functional perspective. This conscious broadcast serves to recruit other processes to be used to select an action with which to deal with the current situation. GWT is a theory of how consciousness functions within cognition. Unconscious contexts influence this competition for consciousness. In GWT, and in its LIDA model, learning requires and follows from attention, and occurs with each conscious broadcast.

Given that GWT is a leading model of human cognition and consciousness, it is valuable to explore whether a computational model of GWT can accommodate higher order mental processes. Three different research teams, lead by Stanislas Dehaene, Murray Shanahan, and Stan Franklin, have developed models for instantiating aspects of GWT computationally. In this paper we focus on the LIDA model developed by Franklin and his team. In doing so, we do not mean to suggest that LIDA, or for that matter computational models of cognition based on GWT, are the only AGI models capable of modeling human-level

decision making. We merely consider LIDA to be a particularly comprehensive model and one that includes features similar to those built into other AGI systems.

The LIDA model describes how an agent tries to make sense of its environment and decides what to do next. An action is selected in every LIDA cognitive cycle (see below), of which there may be five to ten in every second. More complex decisions require deliberation over many such cycles. The challenge for a model of cognition such as LIDA is whether it can truly describe complex higher-order decision making in terms of sequences of bottom up, single cycle action selection.

LIDA and Moral Decision Making

LIDA is a model of human cognition, inspired by findings in cognitive science and neuroscience, that is able to accommodate the messiness and complexity of a hybrid approach to decision making. Our task here is not to substantiate one formal approach to ethics in LIDA. Rather, we'll describe how various influences, such as feelings, rules, and virtues, on ethical decisions might be represented within the mechanisms of the LIDA model. The resulting agent may not be a perfect utilitarian or deontologist, and it may not live up to ethical ideals. A LIDA-based artificial moral agent (AMA) is intended to be a practical solution to a practical problem: how to take into account as much ethically relevant information as possible in the time available to select an action.

Our discussion of moral decision making in LIDA will focus on six areas, most involving several questions.

1. Where are bottom-up propensities and values implemented? How does the agent learn new values and propensities, as well as reinforce or defuse existing values and propensities?

2. How are rules or duties represented in the LIDA model? What activates a rule and brings it to conscious attention? How might some rules be automatized to form unconscious rules-of-thumb (heuristics)?

3. How does the LIDA model transition from a single cycle to the determination that information in consciousness needs to be deliberated upon?

4. What determines the end of a deliberation?

5. How can we implement planning or imagination (the testing out of different scenarios) in LIDA?

6. When a resolution to the challenge has been determined, how might the LIDA model monitor whether that resolution is successful? How might LIDA use this monitoring for further learning?

In the section that follows we describe the LIDA model, its architecture, its antecedents, its relationship to other cognitive architectures, its decision-making, and its learning processes. After that we return to discussing how the LIDA model might be used for moral decision-making. In particular, we offer hypotheses for how the LIDA model answers each of the questions raised in the six issues listed above. Through this exercise we hope to demonstrate the usefulness of a computational model of GWT, and how a computer system might be developed for handling the complexity of human-level decision-making and in particular moral decision-making. Whether a fully functioning LIDA would be judged to demonstrate the moral acumen necessary for moral agency is, however, impossible to determine without actually building and testing the system.

LIDA

The LIDA Model and its Architecture

The LIDA model is a comprehensive, conceptual and computational[2] model covering a large portion of human cognition. In addition to GWT, the model implements and fleshes out a number of psychological and neuropsychological theories including situated cognition (Varela et al. 1991), perceptual symbol systems (Barsalou 1999), working memory (Baddeley and Hitch 1974), memory by affordances[3] (Glenberg 1997), long-term working memory (Ericsson and Kintsch 1995), event segmentation theory (Zacks, *et al.* 2007),and Sloman's H-CogAff (1999). The comprehensive LIDA model includes a broad array of cognitive modules and processes, a database of which, including known possible neural correlates can be found online at http://ccrg.cs.memphis.edu/tutorial/correlates.html.

LIDA is an extension of IDA, an implemented and running software agent that finds new billets for US sailors at the end of their current tour of duty (Franklin, Kelemen, and McCauley, 1998; Franklin and McCauley, 2003). Parts of LIDA are implemented and running. Others are designed and waiting their turn at

[2] Although the LIDA model is only partially implemented, we claim it as a computational model because each of its modules and most of its processes have been designed for implementation.

[3] Gibson (1979) introduced the term *affordance*, meaning that information about the available uses of an object existed in the object itself. We are using it in the sense that the agent can derive such information from the object.

implementation. One can't simply implement LIDA once and for all. Each distinct implementation of the LIDA architecture as a software agent or a robot must be accomplished within a given domain with its own domain and task specific sensors, effectors and motivations. No single LIDA implementation can control different software agents or robots, as each such control structure must be adapted to operate with its own distinct sensors, effectors and motivations. Franklin's research group is currently actively engaged in producing a computational framework for the LIDA architecture that will serve to underlie and facilitate such implementations. But, LIDA is also a work in progress. The conceptual LIDA model is being added to, most recently by the addition of Zacks' event segmentation theory (Zacks *et al.* 2007).

LIDA is a general cognitive architecture that can encompass moral decision making. A full account of the stream of processes by which it does so appears for the first time in this paper. Earlier papers have described various portions of the model and its architecture in some detail (Franklin and Patterson, 2006; Franklin and Ramamurthy, 2006; Franklin, et al., 2007; Friedlander and Franklin, 2008; Negatu, D'Mello, and Franklin, 2007; Ramamurthy, Baars, D'Mello, and Franklin, 2006). However, none has spelled out the entire, multifaceted, decision making process a la LIDA. While its developers hesitate to claim that LIDA is more general or more powerful than other comprehensive cognitive architectures such as SOAR (Laird, Newell, and Rosenbloom, 1987), ACT-R (Anderson, 1990), Clarion (Sun, 2007), etc., they do believe that LIDA will prove to be both a more detailed and a more faithful model of human cognition, including several forms of learning, that incorporates the processes and mechanisms required for moral decision making.

The LIDA Cognitive Cycle

The LIDA model and its ensuing architecture are grounded in the LIDA cognitive cycle. Every autonomous agent (Franklin and Graesser 1997), human, animal, or artificial, must frequently sample (sense) its environment and select an appropriate response (action). Sophisticated agents such as humans process (make sense of) the input from such sampling in order to facilitate their decision making. Neuroscientists call this three part process the action-perception cycle. The agent's "life" can be viewed as consisting of a continual sequence of these cognitive cycles. Each cycle consists of a unit of sensing, of attending and of acting. A cognitive cycle can be thought of as a cognitive "moment." Higher-level cognitive processes are composed of many of these cognitive cycles, each a cognitive "atom."

Just as atoms have inner structure, the LIDA model hypothesizes a rich inner structure for its cognitive cycles (Baars and Franklin 2003, Franklin et al. 2005). During each cognitive cycle the LIDA agent first makes sense of (see below) its current situation as best as it can by updating its representation of both external and internal features of its world. By a competitive process to be described below, it then decides what portion of the represented situation is most in need of

attention. This portion is broadcast, making it the current contents of consciousness, and enabling the agent to choose an appropriate action and execute it.

Figure 1 shows the process in more detail. It starts in the upper left corner and proceeds roughly clockwise.

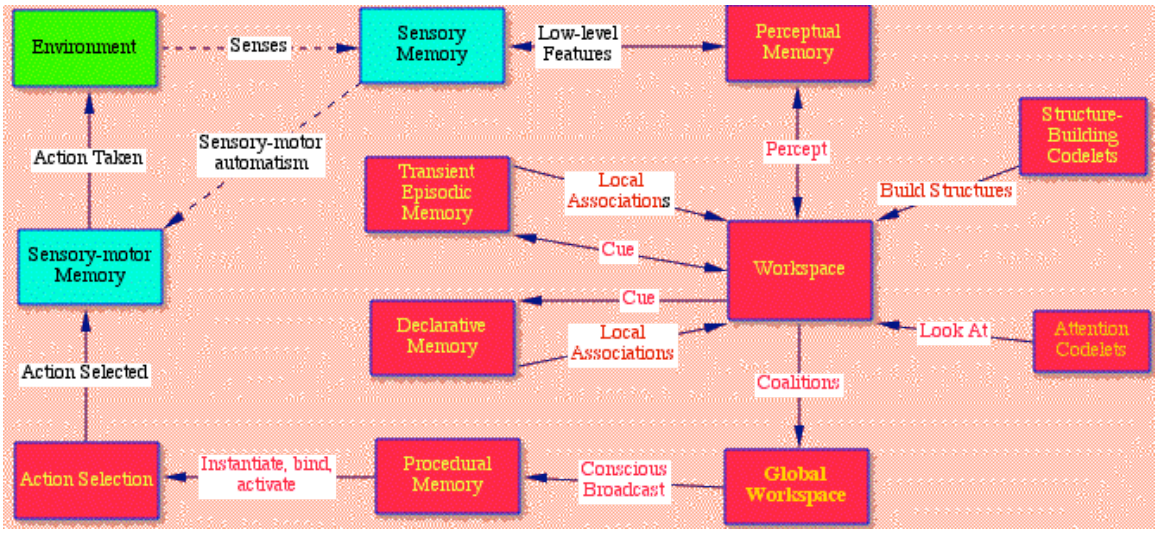

Figure I: LIDA Cognitive Cycle Diagram

The cycle begins with sensory stimuli from external and internal sources in the agent's environment. Low-level feature detectors in sensory memory begin the process of making sense of the incoming stimuli. These low-level features are passed on to perceptual memory where higher-level features, such as objects, categories, relations, situations, etc. are recognized. These entities, which have been recognized preconsciously, make up the percept that passed to the workspace, where a model of the agent's current situation is assembled. This percept serves as a cue to two forms of episodic memory, transient and declarative. Responses to the cue consist of local associations, that is, remembered events from these two memory systems that were associated with the various elements of the cue. In addition to the current percept, the workspace contains recent percepts and the models assembled from them that haven't yet decayed away.

A new model of the agent's current situation is assembled from the percepts, the associations, and the undecayed parts of the previous model. This assembly process will typically be carried out by structure-building codelets[4]. These

[4] The term codelet refers generally to any small, special purpose processor or running piece of computer code. The concept is essentially the same as Baars' processors (1988), Minsky's agents (1985), Jackson's demons (1987), or

structure- building codelets are small, special purpose processors, each of which has some particular type of structure it is designed to build. To fulfill their task these codelets may draw upon perceptual memory and even sensory memory, to enable the recognition of relations and situations. The newly assembled model constitutes the agent's understanding of its current situation within its world. It has made sense of the incoming stimuli.

For an agent operating within a complex, dynamically changing environment, this current model may well be too much for the agent to consider all at once in deciding what to do next. It needs to selectively attend to a portion of the model. Portions of the model compete for attention. These competing portions take the form of coalitions of structures from the model. Such coalitions are formed by attention codelets, whose function is to bring certain structures to consciousness. One of the coalitions wins the competition. In effect, the agent has decided on what to attend.

The purpose of this processing is to help the agent decide what to do next. To this end, a representation of the contents of the winning coalition is broadcast globally, constituting a global workspace (hence the name global workspace theory). Though the contents of this conscious broadcast are available globally, the primary recipient is procedural memory, which stores templates of possible actions including their contexts and possible results. It also stores an activation value for each such template that attempts to measure the likelihood of an action taken within its context producing the expected result. Templates whose contexts intersect sufficiently with the contents of the conscious broadcast instantiate copies of themselves with their variables specified to the current situation. Instantiated templates remaining from previous cycles may also continue to be available. These instantiations are passed to the action selection mechanism, which chooses a single action from one of these instantiations. The chosen action then goes to sensory-motor memory, where it is executed by an appropriate algorithm. The action taken affects the environment, external or internal, and the cycle is complete.

The LIDA model hypothesizes that all human cognitive processing is via a continuing iteration of such cognitive cycles. These cycles occur asynchronously, with each cognitive cycle taking roughly 300 ms. The cycles cascade; that is, several cycles may have different processes running simultaneously in parallel. This cascading must, however respect the serial nature of conscious processing necessary to maintain the stable, coherent image of the world it provides (Merker 2005, Franklin 2005c). Together with the asynchrony, the cascading allows a rate of cycling in humans of five to ten cycles per second. A cognitive "moment" is thus quite short! There is considerable empirical evidence from neuroscience suggestive of and consistent with such

---

Ornstein's small minds (1986). The term was borrowed from Hofstadter and Mitchell (1995).

cognitive cycling in humans (Massimini et al. 2005, Sigman and Dehaene 2006, Uchida, Kepecs and Mainen 2006, Willis and Todorov 2006). None of this evidence is conclusive, however.

Learning in the LIDA model

Edelman (1987) usefully distinguishes two forms of learning, the selectionist and the instructionalist. Selectionist learning requires selection from a redundant repertoire that is typically organized by some form of reinforcement learning. A repertoire of actions is redundant if slightly different actions can lead to roughly the same result. In reinforcement learning (Kaelbling, Littman and Moore 1996) a successfully executed action belonging to an existing repertoire is reinforced, making it more likely to be chosen the next time the result in question is needed. In Edelman's system little-used actions tend to decay away. Instructional learning, in contrast, allows the learning of representations of new actions that are not currently in the repertoire.

Global workspace theory postulates that learning requires only attention (Baars 1988 pp. 213-218). In the LIDA model this implies that learning must occur with each cognitive cycle, because whatever enters consciousness is being attended to. More specifically, learning occurs with the conscious broadcast from the global workspace during each cycle. Learning in the LIDA model follows the tried and true AI principle of generate and test. New representations are learned in a profligate manner (the generation) during each cognitive cycle. Those that are not sufficiently reinforced during subsequent cycles (the test) decay away. Three modes of learning -- perceptual, episodic and procedural -- employing distinct mechanisms (Nadel 1992, Franklin et al. 2005), have been designed and are in various stages of implementation. A fourth, attentional learning, is contemplated but not yet designed. We discuss each individually.

Perceptual learning enables an agent to recognize features, objects, categories relations, and situations. In the LIDA model what is learned perceptually is stored in perceptual memory (Franklin 2005, Franklin 2005a). Motivated by the Slipnet from the Copycat architecture (Hofstadter and Mitchell 1995), the LIDA perceptual memory is implemented as a collection of nodes and links with activation passing between the nodes. Nodes represent features, individuals, categories, actions, feelings and more complex structures. Links, both excitatory and inhibitory, represent relations. Each node and link has both a current and a base-level activation. The base-level activation measures how useful the node or link has been in the past, while the current activation depends on its relevance in the current situation. The percept passed on to the workspace during each cognitive cycle is composed of those nodes and links whose total activation is over the threshold. Perceptual learning in its selectionist form modifies base-level activation, and in its instructionalist form creates new nodes and links. One or the other or both may occur with the conscious broadcast during each cognitive cycle.

Episodic learning refers to the memory of events -- the what, the where and the when (Tulving 1983, Baddeley, Conway and Aggleton 2001). In the LIDA model such learned events are stored in transient episodic memory (Conway 2002, Franklin et al. 2005) and in the longer-term declarative memory (Franklin et al. 2005). Both are implemented using sparse distributed memory (Kanerva 1988), which is both associative and content addressable, and has other characteristics that correspond to psychological properties of memory. In particular it knows when it doesn't know, and exhibits the tip of the tongue phenomenon. Episodic learning in the LIDA model (Ramamurthy, D'Mello and Franklin 2004, 2005) is also a matter of generate and test, with such learning occurring at the conscious broadcast of each cognitive cycle. Episodic learning is initially directed only to transient episodic memory. At a later time and offline, the undecayed contents of transient episodic memory are consolidated (Nadel and Moscovitch 1997, Stickgold and Walker 2005) into declarative memory, where they still may decay away or may last a lifetime.

Procedural learning refers to the learning of new tasks and the improvement of old tasks. In the LIDA model such learning is accomplished in procedural memory (D'Mello et al. 2006), which is implemented via a scheme net motivated by Drescher's schema mechanism (1991). Each scheme in procedural memory is a template for an action, consisting of a context, an action and a result, together with a base-level activation intended to measure how likely the result would be to occur were the action taken within its specific context. Once again, the LIDA model's procedural learning is via a generate and test mechanism, using base-level activation as reinforcement, as well as through the creation of new schemes. These new schemes can support multiple actions, both parallel and sequential.

Attentional learning, that is, the learning of what to attend to (Estes 1993, Vidnyánszky and Sohn 2003) has been relatively little studied by neuroscientists or cognitive scientists (but see Kruschke 2003; Yoshida and Smith 2003). To our knowledge it has been totally ignored by AI researchers, no doubt because few of their systems contain mechanisms for both attention and learning. In the LIDA model attentional learning would involve attention codelets, small processes whose job it is to focus the agent's attention on some particular portion of its internal model of the current situation. When designed, we envision the LIDA model's attentional learning mechanism involving modulating the base-level activation of attention codelets, as well as the creation of new ones.

Feelings and Emotions in the LIDA Model

The word "feeling" may be associated with external haptic sense, such as the feeling in fingertips as they touch the keys while typing. It is also used in connection with internal senses, such as the feeling of thirst, of fear of a truck bearing down, of the pain of a pinprick, of pressure from a full bladder, of shame at having behaved ungraciously, and so on. Here, we are concerned with feelings arising from internal senses.

Following Johnston (1999), in the LIDA model we speak of *emotions* as feelings with cognitive content, such as the joy at the unexpected meeting with a friend, or the embarrassment at having said the wrong thing. The pain in one's arm when scratched by a thorn is a feeling that's not an emotion, because it does not typically involve any cognitive content. Thirst is typically a feeling but not an emotion. Though the boundary between emotions and feelings is fuzzy, the distinction will prove important to our coming discussion of how feelings and emotions motivate low-level action selection and higher-level decision making.

Every autonomous agent must be equipped with primitive motivators, drives that motivate its selection of actions. In humans, in animals, and in the LIDA model, these drives are implemented by feelings (Franklin and Ramamurthy 2006). Such feelings implicitly give rise to values that serve to motivate action selection. Douglas Watt describes well the pervasive role of affect, including feelings, hypothesized by the LIDA model, as seen from the perspective of human neuroscience:

> Taken as a whole, affect seems best conceptualized as a highly composite product of distributed neural systems that together globally organize the representation of value. As such, it probably functions as a master system of reference in the brain, integrating encodings done by the more modular systems supported in various relatively discrete thalamocortical connectivities. Given the central organizing nature of affect as a system for the global representation of value, and given evidence that virtually all stimuli elicit some degree of affective "valence tagging," it would be hard to overestimate the importance of this valence tagging for all kinds of basic operations. The centrality of affective functions is underlined by the intrinsic interpenetration of affect, attentional function, and executive function, and it certainly makes sense that these three global state functions would be highly interdependent. It is logically impossible to separate representation of value from any neural mechanisms that would define attentional foci or that would organize behavioral output. (Watt 1998, p. 114)

Watt's emphasis on "representation of value" and "valence" will be important later for our discussion of the role emotions play in moral decision making. This section will be devoted to an explication of how feelings are represented in the LIDA model, the role they play in attention, and how they act as motivators, implicitly implementing values. (Feelings also act as modulators to learning, as we describe below.) Referring back to the LIDA cognitive cycle diagram in Figure 1 may prove helpful to the reader.

Every feeling has a valence, positive or negative. Also, each feeling must have its own identity; we distinguish between the pains of a pinprick, a burn, or an

insult, and we distinguish pains from other unpleasant feelings, such as nausea. From a computational perspective it makes sense to represent the valence of a single feeling as either positive or negative, that is, as greater or less than zero, even though it may be simplistic to assume that the positive and negative sides of this scale are commensurable. Nevertheless, it may be a viable working hypothesis that in biological creatures feelings typically have only positive valence or negative valence (Heilman 1997). For example, the feeling of distress at having to over-extend holding one's breath at the end of a deep dive is a different feeling from the relief that ensues with the taking of that first breath. Such distress is implemented with varying degrees of negative valence, and the relief with varying positive valence. Each has its own identity. For complex experiences, multiple feelings with different valences may be present simultaneously. For example, the simultaneous fear and exhilaration experienced while on a roller coaster.

Feelings are represented in the LIDA model as nodes in its perceptual memory (Slipnet). Each node constitutes its own very specific identity, for example, distress at not enough oxygen is represented by one node, relief at taking a breath by another. Each feeling node has its own valence, always positive or always negative, with varying degrees. The current activation of the node measures the momentary value of the valence, that is, how positive or how negative. Though feelings are subjected to perceptual learning, their base-level activation would soon become saturated and change very little. Those feeling nodes with sufficient total activations, along with their incoming links and object nodes, become part of the current percept and are passed to the workspace.

Like other workspace structures, feeling nodes help to cue transient and declarative episodic memories. The resulting local associations may also contain feeling nodes associated with memories of past events. These feeling nodes play a major role in assigning activation to coalitions of information to which they belong, helping them to compete for attention. Any feeling nodes that belong to the winning coalition become part of the conscious broadcast, the contents of consciousness. Feeling nodes in the conscious broadcast that also occur in the context of a scheme in procedural memory (the scheme net) add to the current activation of that scheme, increasing the likelihood of it instantiating a copy of itself into the action selection mechanism (the behavior net). It is here that feelings play their first role as implementation of motivation by adding to the likelihood of a particular action being selected. A feeling in the context of a scheme implicitly increases or decreases the value assigned to taking that scheme’s action. A feeling in the conscious broadcast in LIDA also plays a role in modulating the various forms of learning. Up to a point, the higher the affect the greater the learning in the LIDA model. Beyond that point, more affect begins to interfere with learning.

In the action selection mechanism the activation of a particular behavior scheme, and thus its ability to compete for selection and execution, depends upon several

factors. These factors include how well the context specified by the behavior scheme agrees with the current and very recently past contents of consciousness (that is, with the contextualized current situation). The contribution of feeling nodes to the behavior stream's activation constitutes the environmental influence on action selection. As mentioned earlier, the activation of this newly arriving behavior also depends on the presence of feeling nodes in its context and their activation as part of the conscious broadcasts. Thus feelings contribute motivation for taking action to the activation of newly arriving behavior schemes. On the basis of the resulting activation values a single behavior is chosen by the action selection mechanism. The action ensuing from this behavior represents the agent's current *intention* in the sense of Freeman (1999 p. 96ff), that is, what the agent intents to do next. The expected result of that behavior can be said to be the agent's current *goal*. Note that the selection of this behavior was affected by its relevance to the current situation (the environment), the nature and degree of associated feelings (the drives), and its relation to other behaviors, some of these being prerequisite for the behavior.

The selected behavior, including its feelings, is then passed to sensory-motor memory for execution. There the feelings modulate the execution of the action (Zhu and Thagard 2002). Feelings may bias parameters of action such as speed or force. For example, an angry person picking up a soda may squeeze it harder than he would if he weren't angry.

Higher-level Cognitive Processes and Levels of Control

Higher-level cognitive processing in humans includes categorization, deliberation, volition, metacognition, reasoning, planning, problem solving, language comprehension, and language production. In the LIDA model such higher-level processes are distinguished by requiring multiple cognitive cycles for their accomplishment. In LIDA, higher-level cognitive processes can be implemented by one or more behavior streams[5], that is, streams of instantiated schemes and links from procedural memory.

Cognitive processes have differing levels of control. Sloman distinguishes three levels that can be implemented by the architecture of an autonomous agent -- the reactive, the deliberative, and the metacognitive (1999). The first of these, the reactive, is the level that is typically expected of many insects, that is, a relatively direct connection between incoming sensory data and the outgoing actions of effectors. The key point is the relatively direct triggering of an action once the appropriate environmental situation occurs. Though direct, such a connection can be almost arbitrarily intricate, requiring quite complex algorithms to implement in an artificial agent. The reactive level is perhaps best defined by what it's not. "What a purely reactive system cannot do is explicitly construct representations of alternative possible actions, evaluate them and choose between them, all in advance of performing them" (Sloman 1999). Reactive

[5] A stream is a sequence with its order only partially specified.

control alone is particularly suitable for agents occupying relatively simple niches in reasonably stable environments, that is, for agents requiring little flexibility in their action selection. Such purely reactive agents typically require relatively few higher-level, multi-cyclic cognitive processes.

On the other hand, deliberative control typically employs such higher-level cognitive processes as planning, scheduling and problem solving. Such deliberative processes in humans, and in some other animals[6], are typically performed in an internally constructed virtual reality. Such deliberative information processing and decision making allows an agent to function more flexibly within a complicated niche in a complex, dynamic environment. An internal virtual reality for deliberation requires a short-term memory in which temporary structures can be constructed with which to try out possible actions "mentally" without actually executing them. In the LIDA model the workspace serves just such a function. In the earlier IDA software agent, the action selected during almost all cognitive cycles consisted of building or adding to some representational structures in the workspace during the process of some sort of deliberation. Structure-building codelets, the sub-processes that create such structures, modify or compare them, etc., are typically implemented as internal reactive processes. Deliberation builds on reaction. In the LIDA model, deliberation is implemented as a collection of behavior streams, each behavior of which is an internal reactive process (Franklin 2000a). According to the LIDA model, moral decision making will employ such processes.

As deliberation builds on reactions, metacognition typically builds on deliberation. Sometimes described as "thinking about thinking," metacognition in humans and animals (Smith and Washburn 2005) involves monitoring deliberative processes, allocating cognitive resources, and regulating cognitive strategies (Flavell 1979). Metacognition in LIDA will be implemented by a collection of appropriate behavior streams, each with its own metacognitive task. Metacognitive control adds yet another level of flexibility to an agent's decision making, allowing it to function effectively in an even more complex and dynamically changing environmental niche. Metacognition can play an important role in the moral decision making of humans, who may reflect on the assumptions implicit in the values and procedures they apply. However, it would be necessary to implement a fully deliberative architecture before tackling metacognition for any artificial agents, including LIDA.

Deliberation in humans often involves language. Of course metacognition and language have proved to be very difficult challenges for artificial intelligence. While the LIDA model suggests an experimental approach to the challenge posed by language and cognition, detailing that approach is beyond the scope of

[6] Deliberation has been demonstrated in apes (Mulcahy and Call 2006), birds (Werdenich and Huber 2006), and may even be found in arachnids (Wilcox and Jackson 2002, Tarsitano 2006).

this paper. Let it suffice to say that in the conceptual LIDA model, language comprehension is dealt with by word nodes and appropriate links in perceptual memory, leading to structures in the workspace that provide the semantic content of the words. We believe that language generation can be accomplished by schemes in procedural memory whose instantiations produce words or phrases. Given the complexity that language and language creation introduce to the cognitive architecture, the designers of LIDA have tabled this problem until the comprehensive LIDA model has been fully implemented computationally. In the LIDA model, language comprehension is dealt with by word nodes and appropriate links in perceptual memory, leading to structures in the workspace that provide the semantic content of the words. Language generation can be accomplished by schemes in procedural memory whose instantiations produce words or phrases. Given the complexity that language and language creation introduce to the cognitive architecture, the designers of LIDA have tabled this problem until the comprehensive LIDA model has been fully implemented computationally.

Volitional Decision Making

Volitional decision making (volition for short) is a higher-level cognitive process for conscious action selection. To understand volition it must be carefully distinguished from 1) consciously mediated action selection, 2) automatized action selection, 3) alarms, and 4) the execution of actions. Each of the latter three is performed unconsciously. Consciously planning a driving route from a current location to the airport is an example of deliberative, volitional decision making. Choosing to turn left at an appropriate intersection along the route requires information about the identity of the cross street acquired consciously, but the choice itself is most likely made unconsciously -- the choice was consciously mediated even though it was unconsciously made. While driving along a straight road with little traffic, the necessary slight adjustments to the steering wheel are typically automatized actions selected completely unconsciously. They are usually not even consciously mediated, though unconscious sensory input is used in their selection. If a car cuts in front of the driver, often he or she will have turned the steering wheel and pressed the brake simultaneously with becoming conscious of the danger. An alarm mechanism has unconsciously selected appropriate actions in response to the challenge. The actual turning of the steering wheel, how fast, how far, the execution of the action, is also performed unconsciously though with very rapid sensory input.

Though heavily influenced by the conscious broadcast (i.e., the contents of consciousness), action selection during a single cognitive cycle in the LIDA model is not performed consciously. A cognitive cycle is a mostly unconscious process. When speaking, for example, a person usually does not consciously think in advance about the structure and content of the next sentence, and is sometimes even surprised at what comes out. When approaching the intersection in the example above, no conscious thought need be given to the choice to turn left. Consciousness serves to provide information on which such

action selection is based, but the selection itself is done unconsciously after the conscious broadcast (Negatu and Franklin 2002). We refer to this very typical single cycle process as *consciously mediated action selection*.

A runner on an unobstructed sidewalk may only pay attention to it occasionally to be sure it remains safe. Between such moments he or she can attend to the beauty of the fall leaves or the music coming from the iPod. The running itself has become automatized, just as the adjustments to the steering wheel in the example above. In the LIDA model such automatization occurs over time with each stride initiating a process that unconsciously chooses the next. With childhood practice the likelihood of conscious mediation between each stride and the next diminishes. Such automatization in the LIDA model (Negatu, McCauley and Franklin in review) is implemented via pandemonium theory (Jackson 1987).

Sloman (1998) has emphasized the need for an alarm mechanism such as that described in the driving example above. A neuroscientific description of an alarm entails a direct pathway, the “low road,” from the thalamus to the amygdala, bypassing the sensory cortices, the “high road,” and thereby consciousness (Das et al. 2005). The LIDA model implementing alarms via learned perceptual memory alarm structures, bypassing the workspace and consciousness, and passing directly to procedural memory. There the appropriate scheme is instantiated directly into sensory-motor memory, bypassing action selection. This alarm mechanism runs unconsciously in parallel with the current, partly conscious, cognitive cycle.

The modes of action selection discussed above operate over different time scales. Volition may take seconds, or even much, much longer. Consciously mediated actions are selected roughly five to ten times every second, and automatized actions as fast as that, or faster. Alarm mechanisms seem to operate in the sub 50 ms range. In contrast, the execution of an action requires sensory motor communication at roughly forty times a second, all done subconsciously (Goodale and Milner 2004). The possibility of hitting a 90 mph fast ball coming over the plate, or of returning a 140 mph tennis serve makes the need for such sensory motor rates believable.

We now return to a consideration of deliberative, volitional decision making, having distinguished it from other modes of action selection and execution, In 1890, William James introduced his ideomotor theory of volition (1890). James uses an example of getting out of bed on a cold winter morning to effectively illustrate his theory, but in this age of heated homes we will use thirst as an example. James postulated proposers, objectors, and supporters as actors in the drama of acting volitionally. He might have suggested the following scenario in the context of dealing with a feeling of thirst. The idea of drinking orange juice “pops into mind,” propelled to consciousness by a proposer motivated by a feeling of thirst and a liking for orange juice. “No, it's too sweet,” asserts an objector. “How about a beer?” says a different proposer. “Too early in the day,”

says another objector. “Orange juice is more nutritious,” says a supporter. With no further objections, drinking orange juice is volitionally selected.

Baars incorporated ideomotor theory directly into his global workspace theory (1988 chapter 7). The LIDA model fleshes out volitional decision making via ideomotor theory within global workspace theory (Franklin 2000b) as follows. An idea “popping into mind” in the LIDA model is accomplished by the idea being part of the conscious broadcast of a cognitive cycle, that is, part of the contents of consciousness for that cognitive moment. These contents are the information contained within the winning coalition for that cycle. This winning coalition was gathered by some attention codelet. Ultimately, this attention codelet is responsible for the idea “popping into mind.” Thus we implemented the characters in James’ scenario as attention codelets, with some acting as proposers, others as objectors, and others as supporters. In the presence of a thirst node in the workspace, one such attention codelet, a proposer codelet, wants to bring drinking orange juice to mind, that is, to consciousness. Seeing a let’s drink orange juice node in the workspace, another attention codelet, an objector codelet, wants to bring to mind the idea that orange juice is too sweet. Supporter codelets are implemented similarly.

But, how does the conscious thought of “Let’s drink orange juice,” lead to a let’s-drink-orange-juice node in the workspace? Like every higher-order cognitive process in the LIDA model, volition occurs over multiple cycles, and is implemented by a behavior stream in the action selection module. This volitional behavior stream is an instantiation of a volitional scheme in procedural memory. Whenever a proposal node in its context is activated by a proposal in the conscious broadcast, this volitional scheme instantiates itself. The instantiated volitional scheme, the volitional behavior stream, is incorporated into the action selection mechanism, the behavior net. The first behavior in this volitional behavior stream sets up the deliberative process of volitional decision making as specified by ideomotor theory, including writing the let’s-drink-orange-juice node to the workspace[7].

Our fleshing out of ideomotor theory in the LIDA model includes the addition of a timekeeper codelet, created by the first behavior in the volitional behavior stream. The timekeeper starts its timer running as a consequence of a proposal coming to mind. When the timer runs down, the action of the proposal contends in the behavior net to be the next selected action, with the weight (activation) of deliberation supporting it. The proposal is most likely to be selected barring an objection or an intervening crisis. The appearance of an objection in

[7] Alternatively, this node could arrive in the workspace with the percept of the following cycle as a result of internal sensing of the internal speech. In LIDA, this is only an implementation matter, making no functional difference. In humans this is an empirical matter to be decided by experiment. Thus the design decision for LIDA becomes a cognitive hypothesis.

consciousness stops and resets the timer, while that of a supporter or another proposal restarts the timer from a new beginning. Note that a single proposal with no objection can be quickly accepted and acted upon.

But, might this volitional decision making process not oscillate with continuing cycles of proposing and objecting as in Eric Berne's "what if" game (1964)? Indeed it might. The LIDA model includes three means of reducing this likelihood. The activation of a proposer codelet is reduced each time it succeeds in coming to consciousness, thus decreasing the likelihood of its winning during a subsequent cognitive cycle. The same is true of objector and supporter codelets. The LIDA model hypothesizes that supporting arguments help in decision making in part by giving the supported proposal more time in consciousness, allowing more time off the timer. As a second means of preventing oscillation, impatience is built into the timekeeper codelet. Each restart of the timer is for a little less time, thus making a decision easier to reach. Finally, a metacognitive process can watch over the whole volitional procedure, eventually decide that it has gone on long enough, and simply choose an alternative. This latter process has not yet been implemented.

LIDA in Comparison to Other Cognitive Architectures

Competing theories within the cognitive and neuro- sciences suggest different approaches to understanding specific human mental faculties. In describing how the LIDA model handles various tasks, we do not mean to suggest that other approaches are incorrect. However, it is beyond the scope of this paper to discuss the competing theories or approaches. The LIDA model attempts to formulate an approach to AGI that accommodates a significant portion of what is known about human functioning through the work of cognitive scientists and neuroscientists. It is possible that researchers will eventually demonstrate that GWT, upon which the LIDA model has been built, is inadequate for understanding human cognition.

LIDA differs from most other cognitive architectures in several significant ways. Here's a short, selective, but certainly non-exhaustive list.

Most cognitive architectures are either symbolic or connectionist, though some incorporate aspects of both, e.g., Clarion (Sun, 2007) and ACT-R (Anderson 1990). Strictly speaking, LIDA is neither. Though LIDA's internal representations are mostly comprised of nodes and links, the nodes are not symbolic, i.e., amodal. Rather, they should be thought of as perceptual symbols or perceptual symbol generators in the sense of Barsalou (1999). Also, passing activation occurs throughout the LIDA architecture, but none of it is quite in the mode of artificial neural networks. For example, major modes of learning in LIDA are not performed by changing weights on links. Rather, in instructionist learning, new representations are added appropriate to the particular mode, nodes and links to

perceptual associative memory, Boolean vectors to transient episodic memory, or schemes to procedural memory, In selectionist learning the base-level activations of old representations are boosted or diminished.

Following GWT, the LIDA architecture incorporates a specific attention mechanism that selects the most salient, e.g., important, urgent, insistent, portion of its understanding of its current situation for broadcast to all of the modules of the architecture. This broadcast serves both to recruit possible appropriate response actions, and to effect several modes of learning. Other than the much less comprehensive models of Shanahan (2006) and Dehaene's also much less comprehensive neural network model (Dehaene, Sergent, and Changeux, 2003), LIDA is the only such cognitive architecture.

Many of the other general cognitive architectures mentioned above incorporate some form of learning. However, LIDA is unique, to our knowledge, in enabling four distinct modes of learning, perceptual, episodic, procedural and attentional, each modeled after the corresponding mode of human learning. Each mode is human-like in the sense that the learning is unsupervised, continual, and is both selectionist and instructionalist.

Every cognitive architecture must operate via an iteration of sense-cognize-act cycles.[8] The LIDA architecture is unique in distinguishing low level, single cognitive cycle action selection from higher-level multi-cyclic decision making. LIDA's cognitive cycle, hypothesized to occur at a 10hz rate in humans, can be thought of as a cognitive atom or moment, from sequences of which higher-level cognitive processes can be implemented in a consistent fashion.

Though there's been much research on artificial feelings and emotions (e.g., Canamero, 2003; Gadanho, 2003), to our knowledge LIDA is the only comprehensive cognitive architecture to incorporate feelings and emotions as its sole implementation of motivations for action selection, as well as for modulators of learning (Franklin and Ramamurthy, 2006).

[8] This doesn't mean that these cycles need be in strict serial order. Many of the processes within a cycle can operate in parallel. And, the cycles can overlap or cascade. In the LIDA model only the conscious broadcast and the action selection must occur in serial order.

Processing Moral Considerations Within the LIDA Model

Bottom-Up Propensities, Values, and Learning

Complex moral faculties involve reflection about and modification of the bottom-up propensities embodied in emotional/affective responses to actions and their outcomes. Bottom-up propensities in the form of feelings and inherent values influence morality but they are not necessarily reflective of the values a society would recognize as moral values. Negative feelings may, for example, lead to prejudices by automatically attaching to entities that are not a part of the agent's immediate group. From a moral perspective, it is important to understand how top-down considerations interact with these bottom-up propensities reinforcing "good" ones and defusing if not actually eliminating "bad" ones. The approach LIDA offers to the challenge of implementing this hybrid system begins with the way an agent captures bottom-up propensities and the values implicit in these propensities.

Associations between objects, people, contexts, actions, situations, etc. and specific feelings and their valences (positive or negative) are the primary way values and bottom-up propensities form in an agent's mind. The values are implicit in the feelings and their valences, and LIDA captures this dynamic. These associations may arise during perception where sensory input is connected to nodes (objects, feelings, ideas, categories, actions) in perceptual memory. These nodes in turn activate and connect to information retrieved from the various memory systems, which in LIDA are represented as separate memory modules.

Feelings and perceptions that arise within the same LIDA cycle may form associations, particularly when the affective input is strong. But unless the sensory input is particularly strong and sustained, or the initial input cues associated memories, the perception of the objects and situations, and their associated affects, decay quickly and disappear.

The strength of a value, the strength of the connection between feeling and object or situation, is reinforced by sustained sensory input, but these values are short-lived unless the information comes to attention. Attention reinforces a connection for the longer term through perceptual and episodic learning. Powerful memories, that is, memories linked to strong valences, are reinforced each time they come to attention.

LIDA's perceptual memory (a part of long-term memory) is implemented by a Slipnet, a network of nodes, and links between the nodes, that represent structures and concepts. Features, objects, and valenced feelings can be nodes, and links between these nodes represent relationships that can form more complex structures (percepts). These percepts pass on to the system's working

memory[9] (workspace) from which they cue associated information in other areas of short and long-term memory, and this information in turn leads to further associations that may enrich or alter the percepts.

Particularly difficult challenges for LIDA, similar to those encountered by any human-like computer architecture, are how sensory input leads to the activation of nodes in the Slipnet and how new nodes can be created to represent new phenomena. In principle, individual subroutines or codelets can search for specific sensory input, process that information, and pass it on to the activation of a node. Or, similarly, a neural network might organize sensory input. But using current computational technology it becomes difficult to scale either of these approaches to manage a broad array of inputs and nodes.

In addition, there is the difficult problem of determining how to represent valences in the Slipnet. Must they be represented as somatic feelings or is it adequate to use a cognitive representation of the valence. If the feeling is expunged of any somatic affect, and serves merely as a symbol or mathematical formula representing the positive or negative feeling, will it carry the full import of the feeling as it is factored into the selection of an action?

These are not easy problems, but LIDA does offer an architecture for integrating presently available solutions. Given the modularity of LIDA, it will also be able to integrate more sophisticated solutions to these challenges as they emerge from laboratories focusing on the development of specific hardware and software tools.

Moral Deliberation Involving Rules

In almost all situations our action selection decisions, including those that could be said to involve morals, are made in a bottom up fashion during a single cognitive cycle as described above. Much more rarely, but still with some frequency, our moral decisions are more complex and require some thought, that is, deliberation. Such a situation might occur when we are faced with a moral dilemma. This often leads to conflicting voices in our heads, some of which might frame their arguments in terms of rules, for example, “thou shalt not kill”. Let’s consider how such rules are represented in the LIDA model. What activates a rule and brings it to conscious attention? How might some rules be automated to form rules-of-thumb?

A specific example of an inner dialogue about a moral dilemma may help. Suppose the company you work for licenses some new, expensive computer software, say Adobe's *Photoshop*. After becoming comfortable with the new software package at work, you feel the urge to copy it onto your home computer.

[9] Working memory, in the way psychologists use the term (Baddeley 1992) includes consciousness. In the LIDA model, working memory (the workspace) is preconscious in each cognitive cycle.

An internal dialog commences, but not necessarily as wholly verbal and grammatical as what follows. “Let’s bring Photoshop home and load the program on my Mac.” “You shouldn’t do that. That would be illegal and stealing.” “But I’d use it for work related projects that benefit my company, which owns the software.” "Yes, but you’d also use it for personal projects with no relation to the company.” “True, but most of the work would be company related.” Etc., etc., etc.

In such a case, one's decision making is happening consciously, volitionally. The LIDA model describes the handling of such a situation by means of a higher order, multi-cyclic, deliberative process. This conscious, volitional process was described earlier in the section entitled *Volitional Decision Making*. Recall that the internal players included proposers, objectors, supporters, and a timekeeper. Each of the first three players is implemented in the model by an attention codelet that brings ideas to consciousness.

In our example, a proposer, winning the competition for consciousness, causes the idea of copying Photoshop to the home Mac to “pop into mind.” This proposal in consciousness impels the instantiation of a deliberation scheme whose first action, the birth of the timekeeper, the starting of the timer and the writing of the proposal node to the workspace, is selected as the action of the current cycle. In a subsequent cycle that follows soon after, an objector succeeds in bringing to consciousness the idea of “no that would be stealing.” The action selected in this cycle would be to stop and reset the timer. A supporter brings the next idea to consciousness, the timer is restarted, and the process continues over the succeeding cognitive cycles. The game is afoot.

Note that the first objector implicitly based its objection on the rule “thou shalt not steal.” To describe how and where this moral dictum is represented in the LIDA model, and how it plays its role, we begin at the end of the proposal cycle with the proposal structure (“let’s copy *Photoshop*”) in the workspace. There, because of a prior semantic association between copying and stealing it cues the rule “thou shalt not steal” from semantic memory, a part of declarative memory (Franklin et al. 2005). The rule is represented as a structure in the workspace, that is, as a collection of nodes and links from perceptual memory, the common currency for information in the LIDA model (Franklin 2005b). An objector attention codelet then forms a coalition whose informational content is “don’t copy *Photoshop*, that would be stealing.” This objection coming to consciousness and stopping the timer constitutes the rule playing its role in moral decision making.

Rules and duties are stored in semantic memory as perceptual structures. Cued by a proposal or an objection the rule is recalled into working memory as a local association and brought to consciousness to participate in the internal dialogue. Note that a supporter, as well as an objector, can invoke a rule. The dialogue stops when a proposal is on the table without further objection long enough for the timer to ding. At that point a scheme in procedural memory that knows how to

act on the proposal is instantiated into the action selection mechanism with a high activation. Thus its selection is assured barring some crisis or other alarm.

But sometimes this kind of top-down, rule-based decision making can shift to a bottom-up, affect-based action selection. Each time an application of a rule or duty comes to consciousness it, like every conscious event, becomes subject to perceptual learning. If a particular rule is applied frequently in similar situations, LIDA may produce a category node in perceptual memory, representing that rule in an abstract version of the similar situations. In our example, our moral decision making agent might learn the abstract node "don't copy software if you don't have a license for it". If such a node is reinforced often enough this application of the rule is automatic. During the extended learning process the node would acquire links to other nodes, particularly to feelings with negative valence. Thus when faced with a situation where copying software might be tempting, this rule node can become part of the percept. Its presence in the workspace would then inhibit proposer codelets from proposing copying software, that is, by invoking the rule automatically.

Why does the internal dialogue begin? We've seen *how* it begins. It begins with a proposer-attention codelet bringing a proposal into mind, into the global workspace, that is "popping it into mind." But, why isn't the action, copying the software for example, simply selected as the consciously mediated action at the end of a single cycle, with no dialogue at all? In some specific situations copying software is permissible. The software license may allow installation on two machines, office and home, for use by a single user. If encountered frequently enough, a scheme for copying software can be procedurally learned with this situation as its context. In such a case, copying software can become a consciously mediated action that is selected during a single cycle[10]. But, in order for such a scheme to be procedurally learned, its action must have been selected volitionally at least once; that is, some deliberative process must have allowed it.

Generally it's the perceived novelty of a given situation that leads to it being the subject of deliberation, rather than simply selected. It's the newness, or at least apparent newness, of a situation that in effect demands that the agent think about it. New situations do not fit neatly into innate or learned heuristics, and therefore these situations demand attention. In attending to new circumstances, associated proposals and objections naturally come to mind.

[10] The actual process is a little more complex. A behavior stream whose behaviors result in copying software would be instantiated into LIDA's action selection mechanism and the first behavior in that stream likely would be selected.

The Implementation of Planning and Imagination

Moral decision making in humans often involves the planning of various possible scenarios and the testing of them in our imagination. Imagination entails the creation of mental images of objects, actions, situations that are not necessarily current in the outside world. The material for this personal mental realm derives from present and past perceptions of the outside material world and may include some imaginary elements or revisions to existing elements.

The testing of multiple scenarios will, of course, require many cognitive cycles. Some cycles may be devoted to examining an internal scenario while others may entail actions performed on or with external objects. As an example, consider an architect who has been given the task of designing a house for a wooded lot while saving as many trees as possible. Part of the architect's training would have involved learning complex internal behavior streams for constructing and manipulating scenarios by placing various rooms at particular locations. Other internal behavior streams would allow the evaluation of such scenarios (mental floor plans on the lot) using functional, aesthetic, and moral criteria. Volitional decision making, as described above, would employ yet other behavior streams to decide which of the constructed scenarios to select. Appropriately, in LIDA, the central site for much of this work is the workspace, though an embodied LIDA-based robot might also put ideas on paper. This evaluation of possible scenarios could be accomplished without actually cutting down a single tree, and before drawing any building plan. Deliberation will have done its job.

As we've seen previously, each agent that is controlled by LIDA's architecture, including we humans (presuming that LIDA captures the way we function), will understand its environment by means of a model built in the workspace by structure-building codelets. The components of which this internal model of the world is built are nodes and links from perceptual memory, the common currency of the LIDA architecture (Barsalou 1999). The agent's internal representation serves to model both the agent's external environment and its internal environment. We hypothesize this internal representation in the workspace as the site of the structures that enables imagination including deliberation on multiple scenarios.

These internal deliberative structures are built in the workspace using, among other things, material written there over multiple cognitive cycles by behaviors selected at the end of a cognitive cycle whose actions are internal, that is, actions that effect changes within the agent itself, rather than on the outside world. The results of such internal actions may be perceived by the agent through its internal senses, or may be written directly to the workspace, and may also in turn be externalized when, for example, the architect adds a new element to a drawing of the building. All of these possibilities occur in humans.

Ultimately, for moral deliberation to be appropriately modeled by LIDA, attention codelets that are sensitive to morally relevant information will need to be

designed. Whether the design of such morally sensitive codelets differs from the general design of codelets that search for concrete information remains to be seen. But minimally, for example, we expect that attention codelets that are sensitive to concrete information about the facial and vocal expressions of people affected by an AMA's actions will need to be part of the mix. The advantage of codelets is that they provide an indefinitely extensible framework for taking more and more of the relevant factors into account.

The selected internal behaviors that contribute to a deliberation are organized into behavior streams that serve to implement the deliberative process at hand. Such deliberative behavior streams would typically be a product of procedural learning. In our architect's save-the-trees example, a complex behavior stream with behaviors to construct a scenario placing various rooms at particular locations would have been learned. Another internal behavior stream would allow the evaluation of such scenarios (mental or drawn floor plans on the lot) using functional, aesthetic, and moral criteria. Volitional decision making, as described above, would employ yet another behavior stream to decide which of the constructed scenarios to select. Appropriately, the central site for much of this work is the workspace, sometimes complemented by those elements of the deliberation that have been concretized into external forms such as the architects drawing, a mathematical formula, a painting, or a list of criteria by which the scenario should be evaluated.

As described in the section above on *Higher-level Cognitive Processes and Levels of Control,* metacognition in the LIDA model involves the use of deliberation in much the same way as the kind of planning we've just described. Metaethical reflections would be a special case of such metacognition, when the issue at hand was the efficacy or appropriateness of a moral rule or criterion. As mentioned above, we introspectively presume that language and inner voices are central to metaethical reflections. However, the fleshing out of metacognition and metaethics is far beyond the scope of this article, and beyond anything that has been implemented in the LIDA model to date.

Resolution, Evaluation, and Further Learning

A LIDA-based agent would reach a resolution to a volitional decision when there is no longer an objection to a proposal. Given that the activation of an objection decays in repeat cycles, strongly activated proposals will in time prevail over weak objections. However, attention codelets responsible for proposals and their supporters also weaken in their activation as they succeed in coming into consciousness during multiple cycles. Weak proposals may also lose the competition for attention to other concerns demanding attention, defusing any pressure or need for the agent to act on the challenge. Highly activated rules, duties, or other objectors will outlast weak proposers, and force the development of more creative proposals that accommodate the strong objections.

However, time pressures may force a decision before all objections have been dispelled. Decay in the strength of proposals and objections, time pressures on decision making, and pressures from other concerns can drive the selection of a response to a challenge even when the response is inadequate or incomplete. Two mechanisms facilitate dealing with time pressures in the LIDA model. An attention codelet noting the time frame within which the decision must be made would actually increase its activation as the deadline neared. The second mechanism is the timekeeper, discussed above, which manages the volitional decision making process. Recall that impatience is built into the timekeeper. Each restart of the timer is for less time, making a decision easier to reach. An attention codelet reminding of an approaching deadline accelerates this process by continually reducing the time on the timer cycle by cycle.

Furthermore, moral deliberations seldom vanquish all objections even with a generous allocation of time. Moral decisions are often messy, but the LIDA architecture has the means to produce adaptive behavior despite the complexity. Furthermore, future LIDA inspired moral agents may consider a broader array of proposals, objections, and supporting evidence than a human agent can, and thereby, perhaps, select a more satisfactory course of action than many humans.

A LIDA agent, like a human agent, may well be highly susceptible to acting upon strongly reinforced impulses and proposals without necessarily considering the needs of others. That is, the LIDA model in and of itself is morally neutral. What LIDA does offer is a model for computer learning that could provide steps towards a more complete model of moral education.

The manner in which a LIDA-based artificial moral agent monitors its actions will be important to its moral development. When a resolution to a moral challenge has been determined, such an agent monitors the success of the resulting actions as it would any other action, primarily by means of an expectation codelet. An expectation codelet is an attention codelet that is spawned simultaneously with the selected action. The job of this expectation codelet is to bring to consciousness information about the outcome of the action. In particular, the expectation codelet would become strongly activated by discrepancies between the predicted result of a course of action and its actual result. Attention to this discrepancy will in turn reinforce or inhibit the application of that behavior to future similar challenges. In this manner, attention to how the result correlates with the prediction contributes to procedural learning. This general model of procedural learning is applicable to moral development in the context of an agent that has explicitly factored moral considerations into the selection of an action, and into its expectations about the positive moral outcome of

the selected action.

Moving Forward

In this paper we have sought to demonstrate how moral decision making builds upon mechanisms used for the other forms of cognition. LIDA provides one comprehensive model through which to consider the many mechanisms that contribute to the ability to make a moral judgment. Furthermore, we have offered some hypotheses as to how these mechanisms might work together.

The value of a comprehensive theory, such as the GWT/LIDA model, is that it provides a framework for integrating input from a wide variety of sources. A modular system, such as LIDA, can support a broad range of inputs. Modular computer systems do not depend entirely on the ingenuity of one design team. The designers of comprehensive systems can draw on the best-of-breed in the selection of modules developed by other researchers for managing sensory input, perception, or various forms of memory including semantic memory and procedural memory. For example, if a better model than sparse distributed memory became available for transient episodic memory, and that model had been implemented computationally, the new module could be integrated into a LIDA agent instead. The one proviso would be that the output from that module and input to that module could be structured to work with the perceptual nodes in the Slipnet, LIDA's common currency.

In the GWT/LIDA model, competition for consciousness between different coalitions, global broadcasting of the winning coalition, and the selection of an action in each cycle, can be thought of as the mechanisms for integrating the input from the various sources. The unconscious parallel processing of information, the speed of the cycles, and the multi-cyclic approach to higher-order cognitive faculties holds out the promise that a LIDA-like moral agent could integrate a wide array of morally relevant inputs into its choices and actions.

Nevertheless, we do not want to give the impression that AI projects such as LIDA can solve all problems. LIDA, like other AI procedures for choosing actions and testing scenarios, has the problem of scaling--that is a problem of whether its strategy can be adapted to handle the building and evaluation of complex scenarios. Furthermore, the discussion above raises a host of additional questions. Do the mechanisms suggested by these descriptions capture important aspects of the human decision making process? Even if humans function differently, are the mechanisms described adequate to capture the practical demands of moral decision making? Are the mechanisms for representing the conflict between different rules (proposers and objectors) too simplistic to capture the rich

dynamics of human moral decision making? Is the functional model of consciousness suggested by GWT and the LIDA model adequate? Or, will the agent require some form of phenomenal experience that is not captured in the system described? Can morality really be understood without a full description of its social aspects? How well would LIDA handle the kinds of delicate social negotiations that are involved in managing and regulating the conflicts that arise among agents with competing interests?

While we, and others working with the model, are able to suggest ways that LIDA could meet these challenges, initially these approaches will be only theories with no proof of concept. For example, we are aware that LIDA will need something like a Theory of Mind (ToM) to function adequately within social contexts, and are working through ways that the model might be adapted to accommodate an appreciation of other's beliefs and intents. We believe that it may be possible to build a ToM into the model using its existing modules and processes (Friedlander and Franklin. 2008), but as of this writing there is no ToM in LIDA. Certainly the structure-building and attention codelets sensitive to the emotional expressions on people's faces that were mentioned earlier would be an aspect of building a ToM into LIDA.

Of course, many will remain suspicious of mechanical explanations of moral faculties. But the proof, as has been often said, will be in the pudding. What has been described above is certainly not a demonstration that fully functioning AMAs will emerge from computational systems. Rather, we have outlined one rich experimental framework for exploring this possibility.

References

Allen , C. (2002) "Calculated Morality: Ethical Computing in the Limit." In Smit, I. and Lasker, G. (eds.), *Cognitive, Emotive and Ethical Aspects of Decision Making and Human Action,* Vol I, Windsor, Ontario: IIAS.

Allen, C., Smit, I., Wallach, W. (2006) “Artificial Morality: Top-Down, Bottom-Up and Hybrid Approaches”, in *Ethics of New Information Technology,* 7:149-155.

Allen, C., Varner, G., Zinser, J. (2000) “Prolegomena to Any Future Artificial Moral Agent”, *Journal of Experimental and Theoretical Artificial Intelligence,* Vol 12:251-61.

Allen, C. Wendell Wallach, Iva Smit, "Why Machine Ethics*?," IEEE Intelligent Systems*, vol. 21, no. 4, pp. 12-17, Jul/Aug, 2006.

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Anderson M. and Anderson S. (guest editors) Machine Ethics, IEEE Intelligent Systems, Vol 21 Number 4, July/August 2006

Anderson, M., Anderson, S., Armen, C. (2005) “Towards Machine Ethics: Implementing Two Action-Based Ethical Theories”, In Anderson, M., Anderson, S., Armen, C. (eds) *Machine Ethics*, Technical Report FS-05-06, Menlo Park, CA: AAAI Press.
Anderson, M., Anderson, S., Armen, C. (2006) “An Approach to Computing Ethics”, *IEEE Intelligent Systems*, July/August
Baars, B. J. 1988. *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.
Baars, B. J. 2002. The conscious access hypothesis: origins and recent evidence. *Trends in Cognitive Science* 6:47-52.
Baars, B. J., and S. Franklin. 2003. How conscious experience and working memory interact. *Trends in Cognitive Science* 7:166-172.
Baddeley, A. 1992. Consciousness and working memory. *Consciousness and Cognition* 1: 3–6.
Baddeley, A., M. Conway, and J. Aggleton. 2001. *Episodic Memory*. Oxford: Oxford University Press.
Baddeley, A. D., and G. J. Hitch. 1974. Working memory. In The Psychology of Learning and Motivation, ed. G. A. Bower. New York: Academic Press.
Barsalou, L. W. 1999. Perceptual symbol systems. *Behavioral and Brain Sciences* 22:577-609.
Beauchamp and Childress (2001), Principles of Biomedical Ethics, Fifth Edition, Oxford University Press.
Bentham, J. 1961. *An Introduction to the Principles of Morals and Legislation*. Garden City: Doubleday. Originally published in 1789.
Berne, E. 1964. *Games People Play - The Basic Hand Book of Transactional Analysis*. New York: Ballantine Books.
Breazeal, C. 2003. Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies* 59:119-155.
Breazeal, C. (2002) *Designing Sociable Robots*, Cambridge, MA: MIT Press.
Bringsjord, S., Konstantine, A., and Bello, P. (2006) Toward a General Logicist Methodology for Engineering Ethically Correct Robots, *IEEE Intelligent Systems*, July/August
Brooks, R. A. (2002) Flesh and Machines, Pantheon Books, NY
Canamero, Lola D. (2003). Designing Emotions for Activity Selection in Autonomous Agents. In R. Trappl, P. Petta and S. Payr (Eds.), *Emotions in Humans and Artifacts*. Cambridge, MA: MIT Press.
Conway, M. A. 2002. Sensory-perceptual episodic memory and its context: autobiographical memory. In *Episodic Memory*, ed. A. Baddeley, M. Conway, and J. Aggleton. Oxford: Oxford University Press.
Das, P., A. H. Kemp, B. J. Liddell, K. J. Brown, G. Olivieri, A. Peduto, E. Gordon, and L. M. Williams. 2005. Pathways for fear perception: Modulation of amygdala activity by thalamo-cortical systems. *NeuroImage* 26:141-148.
Danielson, P. (1992) Artificial Morality:Virtuous Robots for Virtual Games, New York: Routledge.

Dehaene, S., Sergent, C., and Changeux, J.-P. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proc. Natl. Acad. Sci. USA, 1001*, 8520–8525.
DeMoss, D. (1998). Aristotle, connectionism, and the morally excellent brain, *Proceedings of the 20th World Congress of Philosophy,* The Paideia Archive, URL = http://www.bu.edu/wcp/Papers/Cogn/CognDemo.htm
de Waal, Frans. 2006. *Primates and philosophers: How morality evolved*. Princeton: Princeton University Press.
D'Mello, S. K., U. Ramamurthy, A. Negatu, and S. Franklin. 2006. A Procedural Learning Mechanism for Novel Skill Acquisition. In *Workshop on Motor Development*: *Proceeding of Adaptation in Artificial and Biological Systems, AISB'06*, vol. 1, ed. T. Kovacs, and J. A. R. Marshall. Bristol, England: Society for the Study of Artificial Intelligence and the Simulation of Behaviour; April 2006.
Drescher, G. L. 1991. *Made-Up Minds: A Constructivist Approach to Artificial Intelligence*. Cambridge, MA: MIT Press.
Edelman, G. M. 1987. *Neural Darwinism*. New York: Basic Books Ericsson, K. A., and W. Kintsch. 1995. Long-term working memory. *Psychological Review* 102:211-245.
Estes, William K. 1993. *Classification and cognition*. Oxford: Oxford University Press.
Flavell, J. H. 1979. Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist* 34:906-911.
Franklin, S. 2000a. A "Consciousness" Based Architecture for a Functioning Mind. In *Proceedings of the Symposium on Designing a Functioning Mind, Birmingham, England, April 2000*, ed. A. Sloman.
Franklin, S. 2000b. Deliberation and Voluntary Action in 'Conscious' Software Agents. *Neural Network World* 10:505-521.
Franklin, S. 2003. IDA: A Conscious Artifact? *Journal of Consciousness Studies, 10*, 47–66.
Franklin, S. 2005. Cognitive Robots: Perceptual associative memory and learning. In Proceedings of the 14th Annual International Workshop on Robot and Human Interactive Communication (RO-MAN 2005).
Franklin, S. 2005a. Perceptual Memory and Learning: Recognizing, Categorizing, and Relating. Symposium on Developmental Robotics; American Association for Artificial Intelligence (AAAI); Stanford University, Palo Alto CA, USA. March 21-23, 2005.
Franklin, S. 2005b. Cognitive robots: Perceptual associative memory and learning. In *Proceedings of the 14th annual international workshop on robot and human interactive communication (ro-man 2005)*:427-433.
Franklin, Stan. 2005c. Evolutionary pressures and a stable world for animals and robots: A commentary on Merker. *Consciousness and Cognition* 14: 115–118.
Franklin, S., B. J. Baars, U. Ramamurthy, and M. Ventura. 2005. The Role of Consciousness in Memory. *Brains, Minds and Media* 1:1-38, pdf.

Franklin, S., and A. C. Graesser. 1997. Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. In *Intelligent Agents III*. Berlin: Springer Verlag.

Franklin, S., and Patterson, F. G. J. (2006). The LIDA Architecture: Adding New Modes of Learning to an Intelligent, Autonomous, Software Agent *IDPT-2006 Proceedings (Integrated Design and Process Technology)*: Society for Design and Process Science.

Franklin, S., and U. Ramamurthy. 2006. Motivations, Values and Emotions: Three sides of the same coin. In *Proceedings of the Sixth International Workshop on Epigenetic Robotics*, vol. 128. Paris, France: Lund University Cognitive Studies.

Franklin, S., Ramamurthy, U., D'Mello, S. K., McCauley, L., Negatu, A., L., R. S., et al. (2007, November 9-11). LIDA: A Computational model of global workspace theory and developmental learning. Paper presented at the AAAI Fall Symposium on AI and Consciousness: Theoretical Foundations and Current Approaches, Arlington, VA.

Freeman, W. J. 1999. *How Brains Make Up Their Minds*. London: Weidenfeld and Nicolson General.

Friedlander, David and Stan Franklin. 2008. LIDA and a theory of mind. In *Artificial general intelligence 2008*, ed. Pei Wang, Ben Goertzel and Stan Franklin:137-148. Amsterdam: IOS Press.

Gadanho, S. C. (2003). Learning Behavior-Selection by Emotions and Cognition in a Multi-Goal Robot Task. *Journal of Machine Learning Research 4*, 385-412.

Gazzaniga, Michael (2005) "The Believing Brain" in "The Ethical Brain" Dana Press http://www.press.uchicago.edu/Misc/Chicago/1932594019.html

Gibson, James J. 1979. *The ecological approach to visual perception*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Gips, J. (1991) ``Towards the Ethical Robot'', in Ford, K., Glymour, C., and Hayes, P. *Android Epistemology,* Cambridge, MA: MIT Press, Pp. 243-252.

Glenberg, A. M. 1997. What memory is for. *Behavioral and Brain Sciences* 20:1-19.

Goodale, M. A., and D. Milner. 2004. *Sight Unseen*. Oxford: Oxford University Press.

Grau, C. (2006) " There is No 'I' in 'Robots': Robots and Utilitarianism" *IEEE Intelligent Systems*, July/August

Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M., and Cohen, J.D. (2001). An fMRI investigation of emotional engagement in moral Judgment. Science, Vol. 293, Sept. 14, 2001, 2105-2108

Greene, J. and Haidt, J. (2002) How (and where) does moral judgment work? Trends in Cognitive Sciences, 6(12), 517-523. http://www.wjh.harvard.edu/~jgreene/GreeneWJH/Greene-Haidt-TiCS-02.pdf

Greene, J.D., Nystrom, L.E., Engell, A.D., Darley, J.M., Cohen, J.D. (2004) The neural bases of cognitive conflict and control in moral judgment. Neuron, Vol. 44, 389-400.

Guarini, M. (2006) " Particularism and Classification and Reclassification of Moral Cases", *IEEE Intelligent Systems*, 21 (4); 22-28.
Heilman, K M. 1997. The neurobiology of emotional experience. *J. Neuropsychiatry. Clin. Neurosci.* 9: 439–448.
Hofstadter, D. R., and M. Mitchell. 1995. The Copycat Project: A model of mental fluidity and analogy-making. In *Advances in connectionist and neural computation theory, Vol. 2: logical connections*, ed. K. J. Holyoak, and J. A. Barnden. Norwood N.J.: Ablex.
Holland, Owen (ed.) (2003), Machine Consciousness, Imprint Academic
Jackson, J. V. 1987. Idea for a Mind. *Siggart* Newsletter, 181:23-26.
James, W. 1890. *The Principles of Psychology*. Cambridge, MA: Harvard University Press.
Johnston, V. S. 1999. *Why We Feel:The Science of Human Emotions*. Reading MA: Perseus Books.
Kant, E. (1785) Groundwork of the Metaphysics of Morals
Kaelbling, L. P., M. L. Littman, and A. W. Moore. 1996. Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research* 4:237-285.
Kanerva, P. 1988. *Sparse Distributed Memory*. Cambridge MA: The MIT Press.
Kennedy, C. (2004) "Agents for Trustworthy Ethical Assistance" in Smit, I., Lasker, G. and Wallach, W. (eds), *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, *Vol III.* Windsor, Ontario: IIAS.
Kruschke, John K. (2003) Attention in learning. Current Directions in Psychological Science, 12 (5): 171-175
Laird, John E., J, E., Newell, A., and Rosenbloom, Paul S. P. S. (1987). SOAR: An Architecture for General Intelligence. *Artificial Intelligence, 33*, 1–64.
Massimini, M., F. Ferrarelli, R. Huber, S. K. Esser, H. Singh, and G. Tononi. 2005. Breakdown of Cortical Effective Connectivity During Sleep. *Science* 309:2228-2232.
McDermott, D. (forthcoming) What Matters to a Machine?
McLaren, Bruce (2006) Computational Models of Ethical Reasoning: Challenges, Initial Steps, and Future Directions, IEEE Intelligent Systems
Merker, Bjorn. 2005. The liabilities of mobility: A selection pressure for the transition to consciousness in animal evolution. *Consciousness and Cognition* 14: 89–114.
Mill, J. S. 1998. *Utilitarianism*, edited with an introduction by Roger Crisp. New York: Oxford University Press. Originally published in 1861.
Minsky, M. 1985. *The Society of Mind*. New York: Simon and Schuster.
Mulcahy, N. J., and J. Call. 2006. Apes Save Tools for Future Use. *Science* 312:1038-1040.
Nadel, L. 1992. Multiple memory systems: What and why. *J. Cogn. Neurosci.* 4:179-188.
Nadel, L., and M. Moscovitch. 1997. Memory consolidation, retrograde amnesia and the hippocampal complex. *Curr. Opin. Neurobiol.* 7:217-227.
Negatu, A., D'Mello, Sidney K., and Franklin, S. (2007). Cognitively Inspired Anticipation and Anticipatory Learning Mechanisms for Autonomous Agents.

In M. V. Butz, O. Sigaud, G. Pezzulo and G. O. Baldassarre (Eds.), *Proceedings of the Third Workshop on Anticipatory Behavior in Adaptive Learning Systems (ABiALS 2006)* (pp. 108-127). Rome, Italy: Springer Verlag.

Negatu, A., and S. Franklin. 2002. An action selection mechanism for 'conscious' software agents. *Cognitive Science Quarterly* 2:363-386.

Negatu, A., T. L. McCauley, and S. Franklin. in review. Automatization for Software Agents.

Ornstein, R. 1986. *Multimind*. Boston: Houghton Mifflin.

Picard, R. (1997). *Affective Computing.* Cambridge, MA: MIT Press

Powers, T. (2006) “Prospects for a Kantian Machine”, *IEEE Intelligent Systems,* July/August 2006

Ramamurthy, U., Baars, B. J., D'Mello, Sidney K., and Franklin, S. (2006). LIDA: A Working Model of Cognition. In D. Fum, F. Del Missier and A. Stocco (Eds.), *Proceedings of the 7th International Conference on Cognitive Modeling* (pp. 244–249). Trieste: Edizioni Goliardiche.

Ramamurthy, U., S. K. D'Mello, and S. Franklin. 2004. Modified Sparse Distributed Memory as Transient Episodic Memory for Cognitive Software Agents. In *Proceedings of the International Conference on Systems, Man and Cybernetics*. Piscataway, NJ: IEEE.

Ramamurthy, U., S. K. D'Mello, and S. Franklin. 2005. Role of Consciousness in Episodic Memory Processes. Ninth Conference of the Association for the Scientific Study of Consciousness - ASSC9. Pasadena; June 24 - June 27, 2005.

Scassellati, B. (2001) *Foundations for a Theory of Mind for a Humanoid Robot*, PhD Thesis, Dept of Electrical Engineering and Computer Science, MIT. Available at http://www.ai.mit.edu/projects/lbr/hrg/2001/scassellati-phd.pdf

Shanahan, M. P. (2006). A Cognitive Architecture that Combines Internal Simulation with a Global Workspace. *Consciousness and Cognition, 15*, 433-449.

Sigman, M., and S. Dehaene. 2006. Dynamics of the Central Bottleneck: Dual-Task and Task Uncertainty. *PLoS Biol.* 4.

Sloman, A. 1998. Damasio, Descartes, Alarms and Meta-management. In *Proceedings Symposiumon Cognitive Agents: Modeling Human Cognition*. San Diego: IEEE

Sloman, A. 1999. What Sort of Architecture is Required for a Human-like Agent? In *Foundations of Rational Agency*, ed. M. Wooldridge, and A. S. Rao. Dordrecht, Netherlands: Kluwer Academic Publishers.

Smith, J. D., and D. A. Washburn. 2005. Uncertainty Monitoring and Metacognition by Animals. *Current Directions in Psychological Science* 14:19-24.

Stahl, B.C. (2002) “Can a Computer Adhere to the Categorical Imperative? A Comtemplation of the Limits of Transcendental Ethics in IT” in Smit, I. and Lasker, G. (Eds) *Cognitive, emotive and ethical aspects of decision making in humans and in artificial intelligence, Vol. I* Windsor, Ontario: IIAS.

Stickgold, R., and M. P. Walker. 2005. Memory consolidation and reconsolidation: what is the role of sleep? *Trends Neurosci.* 28:408-415.
Sun, R. (2007). The importance of cognitive architectures: An analysis based on CLARION. *Journal of Experimental and Theoretical Artificial Intelligence, 19*(2), 159-193.
Tarsitano, M. 2006. Route selection by a jumping spider (Portia labiata) during the locomotory phase of a detour. *Animal Behavior* 72:1437-1442.
Tulving, E. 1983. *Elements of episodic memory*. Oxford: Clarendon Press.
Uchida, N., A. Kepecs, and Z. F. Mainen. 2006. Seeing at a glance, smelling in a whiff: rapid forms of perceptual decision making. *Nature Reviews Neuroscience* 7:485-491.
Varela, F. J., E. Thompson, and E. Rosch. 1991. *The Embodied Mind*. Cambridge, MA: MIT Press.
Vidnyánszky, Z., and W. Sohn. 2003. Attentional learning: learning to bias sensory competition [Abstract]. *Journal of Vision* 3:174a.
Wallach, W. (2003)" Robot Morals and Human Ethics", in Smit, I, Lasker, L. and Wallach, W., *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, Vol II, Windsor, Ontario: IIAS.
Wallach, W. (2004) "Artificial Morality: Bounded Rationality, Bounded Morality and Emotions", in Smit, I, Lasker, G. and Wallach, W., *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, Vol III. Windsor, Ontario: IIAS.
Wallach, W. (2008) "Implementing Moral Decision Making Faculties in Computers and Robots", *AI and Society,* Vol. 22, Issue 4, pp. 463-475.
Wallach, W., Allen, C., and Smit, I (2008) "Machine Morality: Bottom-up and Top-down Approaches for Modelling Human Moral Faculties", *AI and Society,* Vol. 22, Issue 4, pp. 565-582.
Wallach, W. and Allen, C. (2009) Moral Machines: Teaching Robots Right From Wrong, New York, Oxford.
Wang, Pei, Ben Goertzel, and Stan Franklin. 2008. *Artificial general intelligence 2008*. Amsterdam: IOS Press.
Watt, D. F. 1998. Affect and the Limbic System: Some Hard Problems. *J. Neuropsychiatry. Clin. Neurosci.* 10:113-116.
Weckert, J. (2005) "Trusting Agents" in *Ethics of New Information Technology: CEPE2005*, Enschede, Holland. Pp. 407-412.
Werdenich, D., and L. Huber. 2006. A case of quick problem solving in birds: string pulling in keas, Nestor notabilis. *Animal Behaviour* 71:855-863.
Wilcox, S., and R. Jackson. 2002. Jumping Spider Tricksters: Deceit, Predation, and Cognition. In *The Cognitive Animal*, ed. M. Bekoff, C. Allen, and G. M. Burghardt. Cambridge, MA: MIT Press.
Willis, J., and A. Todorov. 2006. First Impressions: Making Up Your Mind After a 100-Ms Exposure to a Face. *Psychological Science* 17:592-599.
Yoshida, H. and Smith, L.B. (2003) Known and novel noun extensions: Attention at two levels of abstraction. Child Development, 76 (2) 564-577.
Yudkowsky, E. 2001. What is Friendly AI. Available online at www.kurzweilai.net/meme/frame.html?main=/articles/art0172.html.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., and Reynolds, J. R. (2007). Event Perception: A Mind–Brain Perspective. *Psychological Bulletin, 133*(2), 273–293.
Zhu, J., and P. Thagard. 2002. Emotion and action. *Philosophical Psychology* 15:19-36.