Postprint from International Journal of Machine Consciousness Vol. 3, No. 1 (2011) 177-192

CONSCIOUSNESS AND ETHICS: ARTIFICIALLY CONSCIOUS MORAL AGENTS

WENDELL WALLACH
Interdisciplinary Center for Bioethics, Yale University, P. O. Box 208209, New Haven, Connecticut 06520-8209, USA wendell.wallach@yale.edu

COLIN ALLEN
Cognitive Science Program, Indiana University, 1900E 10th St., Eigenmann 819, Bloomington, Indiana 47406-7512, USA colallen@indiana.edu

STAN FRANKLIN
Department of Computer Science, University of Memphis, Dunn Hall 209, Memphis, Tennessee 38152-3240, USA franklin@memphis.edu

What roles or functions does consciousness fulfill in the making of moral decisions? Will artificial agents capable of making appropriate decisions in morally charged situations require machine consciousness? Should the capacity to make moral decisions be considered an attribute essential for being designated a fully conscious agent? Research on the prospects for developing machines capable of making moral decisions and research on machine consciousness have developed as independent fields of inquiry. Yet there is significant overlap. Both fields are likely to progress through the instantiation of systems with artificial general intelligence (AGI). Certainly special classes of moral decision making will require attributes of consciousness such as being able to empathize with the pain and suffering of others. But in this article we will propose that consciousness also plays a functional role in making most if not all moral decisions. Work by the authors of this article with LIDA, a computational and conceptual model of human cognition, will help illustrate how consciousness can be understood to serve a very broad role in the making of all decisions including moral decisions.

Keywords: Consciousness; ethics; moral decision making; artificial moral agents; LIDA; artificial general intelligence; moral psychology.

Safe, appropriate, and socially condoned responses to morally significant situations can often be programmed into artificial agents. But when confronted with more complex challenges, autonomous systems will require higher-order capabilities in order to select a course of action that will minimize harm and maximize sensitivity to moral considerations. The higher-order capabilities drawn upon will differ from situation to situation. This article will explore when consciousness will be required for machines to arrive at safe, morally appropriate, praiseworthy actions.

The advent of increasingly autonomous systems capable of initiating actions that cause harm to humans and other agents worthy of moral consideration, has given rise to a new field of inquiry variously known as machine ethics (ME), machine morality, artificial morality, computational ethics, robot ethics, and friendly AI. The initial goals of ME are practical, not theoretical. Machine morality extends the traditional engineering concern with safety to domains where the

machines themselves will need to explicitly make moral decisions. When designers and engineers can no longer predict what a system will do as it encounters new situations and new inputs, mechanisms will be required that facilitate the agent's evaluating whether available courses of action are safe, appropriate, and societally or morally acceptable. A system that can make such evaluations is functionally moral [Wallach and Allen, 2009].

The prospect for developing machines that make moral decisions (\moral machines") has stimulated interest in whether theories such as utilitarianism, Kant's categorical imperative, and even Asimov's Laws for Robots might be instantiated computationally. Investigation of strategies for building artificial moral agents (AMAs) has also underscored limitations in the way both philosophers and cognitive scientists approach and understand human moral behavior. Assembling a moral machine from the bottom-up draws attention to a broad array of mechanisms that contribute to the selection of safe, praiseworthy actions. Theoretical research on ME together with recent empirical research on moral psychology by cognitive scientists suggests a need for richer more comprehensive models of moral decision making than presently exist [Wallach and Allen, 2009; Wallach, 2010].

Gips [1991] and Allen et al. [2000] introduced two broad approaches for developing artificial agents capable of making moral decisions — the top-down implementation of a theory of ethics or a bottom-up process (inspired by evolutionary psychology and research on moral development) through which the agent explores courses of actions, is rewarded for behavior that is worthy of praise, and learns. In mapping the prospects for engineering AMAs, Allen, Smit, and Wallach [Allen et al., 2006; Wallach et al., 2008; Wallach and Allen, 2009] elucidated limitations inherent in top-down and bottom-up approaches. They propose that hybrids of both may be necessary for developing sophisticated moral machines. They also note that the capacity to reason about morally relevant information will not be the only capability that AMAs will require in order to arrive at appropriate or laudable courses of action in response to the many types of challenges they are likely to confront. These “suprarational” capabilities and social mechanisms include emotions, a theory of mind, sociability, empathy, an understanding of the semantic content of symbols, an embodied relationship with the environment, and consciousness. Additional capabilities might be added to this list. In other words, moral decisions are seldom the result of just one or two dedicated mechanisms. An array of mechanisms, including those used for general cognitive processes, contribute to determining behavioral responses to ethical challenges.

The moral demands on artificial agents that operate within constrained domains are bounded [Wallach, 2004], and therefore will not require all, or perhaps any, of these suprarational capabilities and social mechanisms. Nor will every context require that AMAs emulate the cognitive faculties that humans use to arrive at morally acceptable actions and behavior. But AMAs that interact with humans in social contexts will need to draw upon an array of suprarational capabilities and mechanisms.

The separate fields of research directed at implementing within AI each of these suprarational faculties are proceeding at their own pace. None of these research trajectories (e.g., affective computing, social robotics, or machine consciousness) are explicitly concerned with building AMAs. Within the field of ME, however, there are many questions as to how moral decisions

will be made by artificial agents with or without emotions, physical embodiment, a theory of mind, an appreciation for social interactions and local customs, or consciousness. How limited or successful, for example, will machines without consciousness be at arriving at morally acceptable courses of action?

Computer scientists pursuing machine consciousness (MC) and artificial general intelligence (AGI) [Wang et al., 2008] have come to appreciate that consciousness itself and general decision making require the integration of input from many sources. The models of cognition they have begun developing accommodate a broad array of inputs and processes. The modular construction of these models means that additional capabilities can be added as needed. How well these modules can be integrated and how effectively such systems can be scaled to navigate complex environments will only be discovered in the process of instantiating the models.

Moral decisions will arguably be among the more difficult challenges AGI systems will need to master. Rather than reinventing the wheel, Wallach and Allen considered whether existing AGI models might be adapted to build AMAs. How might the top-down and bottom-up approaches for designing AMAs fit into the AGI models other scientists were developing? Suprarational capabilities and social mechanisms that AMAs will draw upon are also likely to be essential for other tasks systems with AGI will confront. Which suprarational capabilities had already been accommodated, at least conceptually, within computational models for general intelligence?

As a first step in exploring how a comprehensive model of human cognition might be adapted to make moral decisions, Wallach and Allen teamed with Franklin [Wallach and Allen, 2009; Wallach et al., 2010]. Their joint work with LIDA, the computational and conceptual model of cognition developed by Franklin and his colleagues [Franklin and Patterson, 2006; Ramamurthy et al., 2006; Baars and Franklin, 2007] provides a secondary benefit for studying the relationship between ME and MC. LIDA, and its precursor IDA, are computational and conceptual implementations of Baars' Global Workspace Theory [1988]. In earlier work, Franklin demonstrated how IDA provides a glimpse into a functional role for consciousness in the making of decisions [2003]. The collaboration with Wallach and Allen extended this thesis to a special class of decisions, moral decisions. Together they outlined a comprehensive approach for making moral decisions that uses the same cognitive processes used for making general decisions. That is, moral cognition is supported by domain general cognitive processes, even while some special classes of moral decisions may require additional mechanisms.

In this article we focus upon the functional role of consciousness in the making of moral decisions. While it might be presumed that consciousness is a prerequisite for agents making moral decisions when confronted with complex dilemmas, the exact role(s) consciousness plays has never been fully clarified. We will address four questions:

(a) What is the relationship between an agent being conscious and an agent having a capacity to make moral decisions?

(b) Is consciousness (weak or strong?) essential for an agent to make moral decisions?

(c) What cognitive/computational mechanisms serve both consciousness and moral decision making?

(d) Should the capacity to make moral decisions be considered an attribute essential for being designated a fully conscious agent?

2. Consciousness and Ethics: Rights, Responsibilities, and Moral Judgments

Thought experiments regarding the capabilities a non-human entity would require to be a moral agent have long been a staple for philosophers. Legal theorists have also re°ected upon the criteria for granting rights and/or responsibilities to non-human agents, infants, and the mentally impaired. Indeed, this discussion is foundational for arguments by animal rights advocates that some non-human animals, such as great apes, should be granted certain rights as moral patients, entities toward whom we have moral obligations [Singer, 1977; Regan, 1983].

The foreseeable prospect of artificial agents with abilities that are comparable to and may exceed those of humans, is prompting serious re°ection on criteria for attributing moral agency and granting legal personhood [Calverley, 2005; Chopra and White, 2011]. Legal theorists have argued that since non-biological entities such as corporations are legal"persons", there is, at least in a theoretical sense, nothing in the law that would prohibit this designation for a non-biological agent. Calverley [2005] specifically discusses legal personhood for non-biological systems that are shown to have mental states and other attributes of consciousness.

Steve Torrance has taken the lead in analyzing the relationship between consciousness and ethics in the context of future artificial agents with human-like abilities. Torrance focuses on whether artificial agents can be considered either moral patients or, what he calls,"ethical 'producers' (beings who may have moral obligations and responsibilities themselves)" [Torrance, 2008, p. 499]. Torrance outlines an"organic view" in which — in contrast to non-human animals which might have rights as moral patients — artificial agents (non-biological) which lack sentience or phenomenal awareness would not be"genuine subjects of either moral concern or moral appraisal" [Torrance, 2008, p. 503]. According to Torrance,"only biological organisms have the ability to be genuinely sentient or conscious" [Torrance, 2008, p. 503], a claim he bases on the notion that,"moral thinking, feeling and action arises organically out of the biological history of the human species and perhaps many more primitive species" [Torrance, 2008, p. 502].

Although it is not his focus, Torrance acknowledges that the kinds of capacities mentioned as criteria for designating something as a moral producer might be elements of a theory of moral decision making. The capabilities associated with consciousness that he is most concerned with are the ability to feel pleasure and pain that are central for empathy. He argues that, artificial humanoids"are unlikely to possess sentience and hence will fail to be able to exercise the kind of empathic rationality that is a prerequisite for being a moral agent" [Torrance, 2008, p. 495].

Our focus is not upon whether an artificial agent should be granted moral status, but we agree that the organic view is worthy of serious consideration. However, the possibility of developing synthetic emotions, including pleasure and pain, for artificial (non-biological) agents has been of interest to scientists working in the fields of affective computing and machine consciousness

[Picard, 1997; Franklin and Patterson, 2006; Vallverdú and Casacuberta, 2008; Haikonen, 2009]. It is rather early to evaluate whether existing or future implementations of synthetic emotions will or will not lead to the kind of rich emotional intelligence that might be expected of moral agents. Artificial agents incapable of feeling pain or pleasure or lacking empathy may fail in adequately responding to certain kinds (classes) of moral challenges. But the capacity to empathize is not a prerequisite for responding appropriately to all moral challenges.

Much of the discourse in moral philosophy and in practical ethics would suggest that only decisions about intractable social and personal challenges or decisions in which self-centered interests are transcended for the good of others should be designated moral decisions. In our view moral decision making encompasses a much broader range of choices and actions. Any choice in°uenced by consideration for their effects on others is a moral choice. Given that these kinds of value judgments are implicated in most choices where information is incomplete or of questionable accuracy, or where the consequences of possible courses of action cannot be known in advance, a broad range of choices arguably have moral dimensions.

In recent years there has been considerable research demonstrating that much of moral behavior arises from unconscious judgments. Such research does not rule out conscious deliberation; rather it argues that re°ection is less common than may be otherwise presumed. The social intuitionist model [Haidt, 2001; Haidt and Bjorklund, 2008; Haidt and Joseph, 2007], for example, is generally understood as positing the primacy of emotionally activated intuitions over conscious reasoning as the determinant of most moral behavior. Indeed, the model of a rational agent, and therefore a rational moral actor, has been under assault for more than fifty years [Simon, 1955; Tversky and Kahneman, 1974; Greenwald and Banaji, 1995]. In regards to moral judgments, psychologists and social psychologist have demonstrated experimentally that moral behavior can be altered by priming, by relatively minor changes in the situation, and by additional unconscious or non-conscious in°uences [Isen and Levin, 1972; Darley and Batson, 1973; Haney et al., 1973; Hassin et al., 2006]. But there have also been attempts to re-establish the centrality of reasoning in the making of moral decisions and the role re°ection plays in honing unconscious in°uences on moral judgments [Paxton and Greene, 2010].

Moral philosophers have long held that the ought of ethics is not determined by the is of moral psychology, and certainly not by the unconscious in°uences on moral psychology. But designing or teaching artificial agents to act safely and appropriately may depend on modeling a capacity for reasoned re°ection upon what is known about human psychology. How does one put all this together in an agent? How does one combine bottom-up, reactive psychological processes with top-down, deliberative reasoning?

Building upon the LIDA model of cognition, Wallach et al. [2010] provide a first example of how the top-down analysis and bottom-up psychology might be integrated. This model, in principle, could also accommodate additional suprarational capabilities necessary for making moral decisions in specific domains. For the purposes of this paper, however, we direct our attention to the expanded role for consciousness in the making of moral decisions suggested by the model. We will propose that consciousness plays a functional role in the capacity to make all complex decisions, particularly moral decisions. But before discussing this, let us first introduce

in a very cursory manner the LIDA model of cognition[1] and Global Workspace Theory, which LIDA tries to capture computationally.

3. GWT and LIDA

Bernard Baars' Global Workspace Theory (GWT) [1988] is widely recognized as a high-level theory of the role of consciousness in human cognitive processing with significant support from empirical studies [Baars, 2002]. Three different research teams led by Stanislas Dehaene [Dehaene and Naccache, 2001; Gaillard et al., 2009], Murray Shanahan [2006], and Stan Franklin have developed computational models that instantiate aspects of GWT. LIDA, the model we single out for discussion in this paper, was developed by Franklin and colleagues with input from Baars. LIDA provides a particularly useful model to illustrate a role for consciousness in the making of moral decisions. However, we are not suggesting that LIDA is the only AGI system capable of modeling human-level intelligence. LIDA has many features that are similar to those in other AGI models of cognition.

GWT is a neuropsychological model of consciousness that views the nervous system as a distributed parallel system incorporating many different specialized processes. Various coalitions of these specialized processes facilitate making sense of the sensory data currently coming in from the environment. Other coalitions sort through the results of this initial processing and pick out items requiring further attention. In the competition for attention a winner emerges and occupies what Baars calls the global workspace, the winning contents of which are presumed to be at least functionally conscious. The presence of a predator, enemy, or imminent danger should be expected, for example, to win the competition for attention. However, an unexpected loud noise might well usurp consciousness momentarily even in one of these situations. The contents of the workspace are broadcast to processes throughout the nervous system in order to recruit an action or response to this salient aspect of the current situation. The contents of this global broadcast enable each of several modes of learning.

LIDA is a computational agent that continually tries to make sense of its environment and determines what to do next. In LIDA this dynamic is represented through a model that describes how unconscious mechanisms feed the conscious processing of information. LIDA implements GWT as a cascading sequence of cognitive cycles (see Fig. 1).

During a cognitive cycle the agent constantly samples (senses) its external and internal environment. This sensory information is matched to information within various memory systems in a process of discerning the features (objects, categories, relations, events, situations, etc.) of the present situation.

Fig. 1. The LIDA cognitive cycle.

---

[1] LIDA is a very extensive model of cognition, whose development is covered in more than fifty published articles. Many of the articles focus upon facets of LIDA. The most comprehensive description of decision making in LIDA is contained in Wallach et al. [2010] upon which this paper builds.

In each cognitive cycle, attention codelets search this unconscious model of the present situation for specific features or percepts. Thousands of these codelets could be searching specifically for morally relevant considerations. Those that find information germane to their directive or function will join in coalitions with other attention codelets that have found related information. These coalitions occupy the global workspace and vie for attention. In each cycle there is a winning coalition whose contents are made conscious by being broadcast throughout the system. The broadcast is directed in particular at procedural memory in pursuit of an action in response to the information requiring attention. Franklin hypothesizes that there are 5–10 cycles each second [Madl et al., in press]. In each of these 5–10 cycles there is a broadcast and the selection of an action in response.

One might, for example, presume that while driving a car focusing upon the road conditions wins the competition for attention in many cycles, but once the response (or lack of a need for any change) has been determined, the"check the road" coalition will be weaker than other coalitions during a number of intermediary cycles. Attention can now be devoted to listening to the radio during intermediary cycles.

In LIDA a distinction is made between the conscious mediation of the contents of the global workspace leading to the activation of an unconsciously chosen response, and the conscious process of volitional decision making. Generally the former is said to occur in one cognitive cycle, while the latter will require multiple cycles.

4. Moral Judgments and Moral Decisions

Usually decision making is thought of as a deliberative process that entails reasoning, planning, problem solving, and meta-cognition. But action selection can also occur through consciously mediated emotionally activated intuitions and automatized learned behavior. We follow Haidt's nomenclature by referring to emotionally activated intuitions or the unconscious selection of consciously mediated actions as judgments. Judgments do not necessarily express social norms and may even be prejudices, such as a rejection of those who do not belong to one's group. While there is no deliberation or volitional decision making in judgments, future judgments can be altered through experience. Baars has proposed that attention is sufficient for learning and thus learning occurs with each conscious broadcast [1988]. Several modes of such learning have been included in the LIDA model [Franklin and Patterson, 2006]. The bottom-up propensities embodied in emotional/affective responses to actions and their outcomes can also be modified by ex post facto re°ection. In other words, there is an ongoing process where consciously mediated behavior is molded and honed by experience and deliberation. Given this dynamic relationship between judgments and deliberations, the moral decision-making system of an agent can be said to include both.

A strength of LIDA lies in the model's ability to accommodate the messiness and complexity of a hybrid approach to decision making. Moral decisions are particularly messy, drawing upon emotions, moral sentiments, intuitions, heuristics, rules and duties, principles, and even some explicit valuation of utility or expected outcomes. In analyzing how a moral decision-making system that includes both judgments and deliberation might be implemented in LIDA, Wallach et

al. [2010] addressed six questions, whose answers we very brie°y summarize here.[2] In answering these questions it was not our intent to explain how a particular theory of ethics would be implemented in LIDA. Rather, we wish to provide tentative explanations of general processes central to moral judgments and volitional decision making.

(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?

Objects, people, contexts, events, situations, and other percepts are represented as nodes within the perceptual memory of LIDA. Associations between these percepts and valenced feelings (either negative or positive) are the primary means of capturing values within the agent's mind [Franklin and Patterson, 2006]. These values often represent propensities that have been evolutionarily acquired and then shaped by experience.

Affects and perceptions that arise within one LIDA cycle automatically form associations, which will decay over time, but can also be strengthened or weakened by sustained sensory input and attention. As we mentioned earlier, Baars posits that each instance of attention contributes toward learning. By reinforcing links in an association, sustained or conscious attention produces learning and can build longterm memory.

A particularly difficult challenge for all human-like computer architectures is how the system generates or acquires totally new nodes. There is also the difficult problem of how valenced feelings will be represented. A cognitive representation of feelings may not adequately express the richness of content and meaning that we associate with somatic feelings.

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

Generally, when the situation is either new or appears to be novel, a multicycled process of deliberation will naturally arise because LIDA is unlikely during a single cycle to determine an appropriate response. In effect, new or apparently new situations that do not fit neatly into learned heuristics demand attention over time.

(3) 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)?

Deliberation in LIDA is modeled upon William James theory of"volitional" decision making. James viewed decisions as a negotiation between internal proposers of courses of action, objectors, and supporters. In LIDA proposed courses of action that win the competition for consciousness will impel the initiation of an action stream that will continue to be reviewed in

[2] Chapter 11 of Moral Machines: Teaching Robotics Right from Wrong [Wallach and Allen, 2009] provides a fuller discussion, along with examples, of the answers to these questions. The fullest discussion of the answers can be found in Wallach et al. [2010].

subsequent cycles. This action stream could activate a rule or duty, stored in semantic memory, which objects to the proposed action, and in a subsequent cycle wins the competition for attention. Thus begins a multicycled process in which proposer, objector, and supporter codelets successively win the competition for attention. Pertinent memories and other associated information also enter the mix, including additional rules or duties and relevant exceptions.

The activation of proposals and objections decays in each subsequent cycle. But strong constraints, such as not killing, can have high levels of activation through reinforced connections to feelings such as shame or fear. Thus strong constraints are not easily over-ridden. While this kind of process is compatible with a battle of urges that might not be considered a deliberation, when the competing elements correspond to reasons then we believe this is an implementation of a deliberation.

On each occasion when a rule or duty comes to attention it is a subject for learning and modification. For rules that arise in similar situations, or for situations that required sustained deliberation, LIDA will produce a new node in perceptual memory, that might represent a new variation of the rule. If encountered often enough, a response to a challenge for which there is no objection can be captured within procedural memory as an habituated response and activated in just one cognitive cycle. But before this kind of procedural learning can occur, the action must have been selected at least once after a deliberative process. Deliberation can produce habitual behaviors.

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

Imaginative planning or testing of various possible scenarios in LIDA is comparable to building a model in the workspace of the current situation by linking nodes from perceptual and/or episodic memory. The consciously proposed action at the completion of each model or scenario will or will not cue objections indicating the success or issues with the model. In the following cycle, the action selected in response to an objection could be as simple as making a minor alteration or adding a component to the model.

(5) What determines the end of a deliberation?

The deliberative dialogue continues until there is no objection to a proposed action or until the metaphorical timer requiring some action rings. The decay of objections in succeeding cycles means that strongly reinforced proposals will win out over weak objections. Strong objections will prevail over weak proposals. However, time pressures can force a decision before all objections have been dispelled.

The selected course of action may in retrospect fail to satisfy prevailing exterior norms. LIDA-inspired moral agents are not designed around fixed moral rules, values, or principles. Therefore they, like human agents, are prone to acting upon strong impulses without necessarily taking into consideration the needs of others. Developing a LIDA agent with the kind of rich sensitivity to moral considerations and how each consideration might be comparatively weighted will require

much experience and learning, just as it does with teaching a young adult to be morally sensitive and considerate.

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

Monitoring actions is central for moral development. Once a resolution has been reached and an action selected, LIDA generates an expectation codelet, an attention codelet that brings to attention the outcome of an action. An expectation codelet would bring to the global workspace any discrepancy between the actual results of an action and the predicted results. Any discrepancy would contribute to procedural learning inhibiting or reinforcing the application of that procedure to future similar challenges. For a decision that entailed moral considerations, such procedural learning would foster moral development through expectations in regard to the positive or negative moral outcome of actions.

4.1. The advantage of codelets

Attention codelets sensitive to morally relevant information would need to be designed for LIDA to engage in moral deliberations. It is unclear at this time whether the design of such codelets would differ significantly from the design of codelets that search for concrete information. We expect that attention codelets sensitive to facial and vocal expressions would be among those gathering information relevant to moral deliberations.

An advantage of codelets is that they provide an extensible framework through which more and more relevant considerations can be found within the model of the situation. No one needs to specify in advance the considerations in°uencing the situation or the criteria for evaluating a scenario. Therefore, codelets provide a particularly useful approach for representing the messiness of so many moral challenges.

5. Evolution, Moral Decisions and Consciousness

Certainly artificial agents that lack consciousness can make moral decisions in many situations by applying normative rules to the information at hand. But in our analysis consciousness becomes much more central for agents operating within complex environments. These agents will perform many tasks related to the making of moral decisions in addition to applying norms, rules, or moral principles. AMAs will also need to recognize when they are in an ethically significant situation, to have sensitivity to the array of moral considerations that impinge upon that decision, to discriminate essential from inessential information, to estimate the sufficiency of initial information and search for additional information when needed, and to make judgments about the intentionality of the other agents with whom it is interacting. All of this must be done within the time available to take action.

An array of sensors, subsystems, and processes will be enlisted to perform these tasks. These are largely the same cognitive mechanisms used for general decision making. Moral cognition is supported by domain general processes, many of which will also be needed for machine consciousness. So how closely entangled are the fields of machine ethics and machine

consciousness? Should we consider the capacity to make moral decisions an attribute essential to consciousness?

Scholars point to evidence of proto-morality [Katz, 2000; de Waal, 2006; Bekoff and Pierce, 2009] and pre-re°ective consciousness [Seth et al., 2005; Panksepp, 2005; Edelman and Seth, 2009] in many non-human species. The evolution of consciousness and the evolution of morality are each developing as sub-disciplines in their own right. Wallach [2005] hypothesized that consciousness and moral decision making coevolved. The basic argument is that one fitness function of consciousness lay in how it facilitated the making of decisions in ambiguous situations where neither of two or more instinctual or already learned action sequences (automatized responses) was activated. Consider the distant presence of a potentially threatening natural enemy that does not immediately activate a fight or °ight response. Such situations can lead to sustained attention (proto-re°ection) during which existing action sequences are altered or new action sequences are created. The latter might entail the recognition by a mammalian or earlier ancestor that under certain circumstances a natural enemy is non-threatening, for example, the predator is satiated or ready for a nap. The observing agents can thus conserve energy and is free to return to other tasks such as feeding the young, while continuing to periodically check whether her enemy is taking a more threatening stance. This requires a moral decision in the sense that some form of"valuing" would come into responding to the challenge. In other words, under circumstances where information is unclear or incomplete sustained attention (proto-re °ection) leads to a kind of valuing of information that alters existing behavioral patterns, but more importantly, can even create new behavior streams.

If Wallach is correct, then central among the fitness functions served by consciousness is the manner in which it facilitates a kind of valuing that fostered making decisions when the available information is ambiguous, confusing, contradictory, or incomplete. Given that such situations are common, this kind of valuing opened up an array of new possibilities.

Even if one holds that the evolution of consciousness, decision making, and moral decision making are entangled, there are practical reasons for distinguishing between the project of designing AMAs and exploring ways to implement consciousness in machines. However, we should not lose sight of how both projects will progress with the emulation of the same cognitive processes utilizing many of the same cognitive mechanisms. Furthermore, we should not overlook evolutionary strategies that further the interactive development of MC and the design of AMAs.

6. Conscious Artificial Moral Agents

Our discussion of the LIDA model of cognition illustrates a functional role for consciousness in the making of all decisions, including moral decisions. However, neither LIDA nor any other AGI has been fully implemented. Therefore, it is impossible to know whether such systems can adequately be scaled to manage the on-going sorting through sensory information, action selection, scenario building, and deliberation necessary to operate successfully in complex environments filled with other agents. In the process of building models of AGI, designers and engineers will recognize the need for additional subsystems to support capabilities whose contribution may have not yet been recognized. The value of models such as LIDA is that they

offer an architecture that in theory can integrate a vast array of inputs into an approach for discerning salient information, for selecting appropriate actions, and for rich deliberation. Competition for consciousness between different coalitions, global broadcasting of the winning coalition, and the selection of an action in each cycle are the mechanisms used in the LIDA model to integrate input from various sources.

While we cannot know for sure that LIDA-like agents will accommodate a broad array of moral considerations in choices and actions, the multicyclical approach to higher-order cognition, the increasing speed of computers and therefore the speed of individual cycles, and the unconscious parallel processing of information all suggest that this is a promising path to explore in the development of sophisticated AMAs.

We do not know how far this strategy will progress. Nor do we know how broad an array of ethical challenges can be managed successfully within the functional model of consciousness suggested by GWT and LIDA. A further challenge concerns the necessity for some form of phenomenal experience that might not be captured in the GWT/LIDA model. Phenomenal consciousness is difficult to measure in any agent. But if, as Torrance notes in his explication of the organic view, phenomenal awareness is foundational for empathetic rationality, then evidence that an agent is sensitive to what others feel, might be counted as evidence the agent is phenomenally conscious. However, if the organic view is correct and only biological entities can be moral agents, silicon-based artificial agents will never progress beyond a kind of deductive empathy.

In this article, we explored the relationship between consciousness and moral decision making within one model capable of implementing both. Given this implementation, we propose the following:

(1) Consciousness serves most if not all moral decision making as it serves decision making more generally. Consciousness will be especially important for making volitional moral decisions.

(2) Moral cognition is supported by domain general cognitive processes. (3) Among the mechanisms that serve both moral decision making and consciousness are those which integrate perception and action over multiple cycles, including inner modeling of the relationship between the agent and its environment for generating expectations of events and assessing the consequences of various courses of action.

Machine consciousness and designing artificial moral agents are two projects that are joined at the hip. Although we have not argued for it directly, we further propose that the capacity to make moral decisions be considered an attribute essential for being designated a fully conscious agent. Thus, a fully conscious machine or an AGI system worthy of being considered a full moral agent would be an artificial conscious moral agent.

References

Allen, C., Varner, G. and Zinser, J. [2000] "Prolegomena to any future artificial moral agent," Journal of Experimental and Theoretical Artificial Intelligence 12, 251–261.

Allen, C., Smit, I. and Wallach, W. [2006] "Artificial morality: Top-down, bottom-up and hybrid approaches," Ethics of New Information Technology 7, 149–155.

Baars, B. J. [1988] A Cognitive Theory of Consciousness (Cambridge University Press).

Baars, B. J. and Franklin, S. [2007] "An architectural model of conscious and unconscious brain functions: Global Workspace Theory and IDA," Neural Networks 20, 955–961.

Bekoff, M. and Pierce, J. [2009] Wild Justice: The Moral Lives of Animals (University of Chicago Press).

Calverley, D. [2005] "Towards a method for determining the legal status of a conscious machine," in Artificial Intelligence and the Simulation of Behavior '05: Social Intelligence and Interaction in Animals, Robots and Agents: Symposium on Next Generation Approaches to Machine Consciousness, Hatfield, UK.

Chopra, S. and White, L. F. [2011] A Legal Theory for Autonomous Artificial Agents (University of Michigan Press).

Darley, J. M. and Batson, C. D. [1973] "From Jerusalem to Jericho: A study of situational and dispositional variables in helping behavior," Journal of Personality and Social Psychology 27, 100–108.

Dehaene, S. and Naccache, L. [2001] "Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework," Cognition 79, 1–37.

de Waal, F. [2006] Primates and Philosophers: How Morality Evolved (Princeton University Press).

Edelman, D. B. and Seth, A. K. [2009] "Animal consciousness: A synthetic approach," Trends in Neurosciences 32(9), 476–484.

Franklin, S. [2003] "IDA: A conscious artefact?" Journal of Consciousness Studies, 10(4–5), 47–66.

Franklin, S., Baars, B. J., Ramamurthy, U. and Ventura, M. [2005] "The role of consciousness in memory," Brains, Minds and Media 1, 1–38.

Franklin, S. and Patterson, F. G. J. [2006] "The LIDA architecture: Adding new modes of learning to an intelligent, autonomous, software agent," in IDPT-2006 Proceedings (Integrated Design and Process Technology), Society for Design and Process Science.

Franklin, S. and Ramamurthy, U. [2006] "Motivations, values and emotions: Three sides of the same coin," in Proceedings of the 6th International Workshop on Epigenetic Robotics, Paris, France, pp. 41–48.

Gaillard, R., Dehaene, S., Adam, C., Clemenceau, S., Hasboun, D., Baulac, M., Cohen, L. and Naccache, L. [2009] "Converging intracranial markers of conscious access," Consciousness and Ethics: Artificially Conscious Moral Agents, PLoS Biology 7(3), http://www.plosbiology.org/article/info%3Adoi%2F10.1371%2Fjournal.pbio.1000061.

Gips, J. [1991] "Towards the ethical robot," in Android Epistemology, Ford, K. G., Glymour, C. and Hayes, P. J. (eds.) (MIT Press), pp. 243–252. Greenwald, A. G. and Banaji, M. R. [1995] "Implicit social cognition: Attitudes, self-esteem,

and stereotypes," Psychological Review 102, 4–27. Haidt, J. [2001] "The emotional dog and its rational tail: A social intuitionist approach to moral judgment," Psychological Review 108(4), 814–834.

Haidt, J. and Joseph, C. [2007] "The moral mind: How five sets of innate moral intuitions guide the development of many culture-specific virtues, and perhaps even modules," in The Innate Mind, Carruthers, P., Laurence, S. and Stich, S. (eds.) (Oxford University Press), pp. 367–391.

Haidt, J. and Bjorklund, F. [2008] "Social intuitionists answer six questions about moral psychology," in Moral Psychology, Vol. 2: The Cognitive Science of Morality: Intuition and Diversity, Sinnott-Armstrong, W. (ed.) (MIT Press), pp. 181–217.

Haikonen, P. O. A. [2009] "Qualia and conscious machines," International Journal of Machine Consciousness 1(2), 225–234.

Haney, C., Banks, W. and Zimbardo, P. [1973] "Interpersonal dynamics of a simulated prison," International Journal of Criminology and Penology 1, 69–97.

Hassin, R., Uleman, J. and Bargh, J. (eds.) [2006] The New Unconscious (Oxford University Press).

Isen, A. M. and Levin, P. F. [1972] "Effect of feeling good on helping: Cookies and kindness," Journal of Personality and Social Psychology 21, 384–388.

Katz, L. D. (ed.) [2000] Evolutionary Origins of Morality: Cross-Disciplinary Perspectives (Imprint Academic).

Madl, T., Baars, B. J. and Franklin, S. [in press] "The timing of the cognitive cycle," PLoS ONE.

Panksepp, J. [2005] "Affective consciousness: Core emotional feelings in animals and humans," Consciousness and Cognition 14, 30–80.

Paxton, J. and Greene, J. [2010] “Moral reasoning: Hints and allegations,” Topics in Cognitive Science 2, 511–527.

Picard, R. [1997] Affective Computing (MIT Press). Ramamurthy, U., Baars, B. J., D'Mello, Sidney, K. and Franklin, S. [2006] “LIDA: A working model of cognition,” in Proceedings of the 7th International Conference on Cognitive Modeling, pp. 244–249.

Regan, T. [1983] The Case for Animal Rights (University of California Press).

Seth, A. K., Baars, B. J. and Edelman, D. B. [2005] “Criteria for consciousness in humans and other mammals,” Consciousness and Cognition 14, 119–139.

Shanahan, M. P. [2006] “A cognitive architecture that combines internal simulation with a global workspace,” Consciousness and Cognition 15, 433–449.

Simon, H. [1955] “A behavioral model of rational choice,” Quarterly Journal of Economics 69, 99–118.

Singer, P. [1977] Animal Liberation (Granada). Torrance, S. [2008] “Ethics, consciousness and artificial agents,” AI and Society 22(4), 495–521.

Tversky, A. and Kahneman, D. [1974] “Judgment under uncertainty: Heuristics and biases," Science 185, 1124–1131.

Vallverdú, J. and Casacuberta, D. [2008] “The panic room: On synthetic emotions,” in Proceedings of the 2008 Conference on Current Issues in Computing and Philosophy, pp. 103–115.

Wallach, W. [2004] “Artificial morality: Bounded rationality, bounded morality and emotions,” in 16th International Conference on Systems Research, Informatics and Cybernetics: Symposium on Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence, Baden-Baden Germany, International Institute for Advanced Studies in Systems Research and Cybernetics, pp. 1–6.

Wallach, W. [2005] “Choice, ethics, and the evolutionary function of consciousness,” in 17th International Conference on Systems Research, Informatics and Cybernetics: Symposium on Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence, Baden-Baden Germany, International Institute for Advanced Studies in Systems Research and Cybernetics, pp. 1–21.

Wallach, W. [2010] “Robot minds and human ethics: The need for a comprehensive model of moral decision making,” Ethics and Information Technology 12(3), 243–250.

Wallach, W., Allen, C. and Smit, I. [2008] “Machine morality: Bottom-up and top-down approaches for modelling human moral faculties,” AI and Society 22(4), 565–582.

Wallach, W. and Allen, C. [2009] Moral Machines: Teaching Robots Right from Wrong (Oxford University Press).

Wallach, W., Franklin, S. and Allen, C. [2010] "A conceptual and computational model of decision making in human and artificial agents," Topics in Cognitive Science 2, 454–485.

Wang, P., Goertzel, B. and Franklin, S. [2008] Artificial General Intelligence 2008 (IOS Press).