# Exploring Moral Reasoning in a Cognitive Architecture

**Wayne Iba**
Computer Science Department
Westmont College, 955 La Paz Road
Santa Barbara, CA 93108

**Pat Langley**
Institute for the Study of Learning and Expertise
2164 Staunton Court, Palo Alto, CA 94306

### Abstract

Moral reasoning plays a significant but poorly understood role in human action and interaction. Although studied by philosophers for millennia, considerable confusion surrounds the topic. Computational cognitive architectures hold promise for shedding insight on how agents act and reason morally. We present a view of moral cognition and examine one implementation of that view in ICARUS, a theory of the human cognitive architecture. This approach to moral behavior and reasoning leads us to suggest that morality is a special case of everyday cognition. We discuss the implications of this view and outline our continuing research on these and related questions.

**Keywords:** Machine ethics; Moral reasoning; Philosophy; Cognitive architectures.

## Introduction

Moral reasoning has been a focus of study for millennia. The human condition stems in large part from the collective judgments and decisions that can be said to have moral valence. This suggests that it would be desirable to study this topic carefully. Nevertheless, our understanding of morally charged cognition is still incomplete and uncertain. Since the introduction of computers, many people have been fascinated with the prospect of writing programs that exhibit human intelligence. As with other aspects of intelligence, computational models can provide many insights into the nature of moral cognition.

Although generally treated as a topic distinct from generic reasoning, we suggest here that moral cognition is better viewed as a special case of practical cognition. Depending on the ethical approach one takes, moral dilemmas might be nothing other than the consequence of bounded rationality. Perhaps our subjective experience when deliberating over a moral dilemma has more to do with an emotion than with the properties of the problem that we are trying solve. These two factors – bounded rationality and emotional states – may be the actual source of moral dilemmas.

We start by distinguishing between three types of moral cognition: moral behavior, moral interpretation, and moral decision making. Next we review ICARUS, the cognitive architecture we have used to implement agents that exhibit forms of moral cognition. We discuss our experiences with these agents and their broader implications for accounts of this class of phenomena. In closing, we suggest directions for future research that should further clarify our understanding of moral cognition.

## Background

In this paper, we describe our attempt to model moral cognition within an integrated cognitive architecture. As such, we draw upon work in three distinct disciplines – philosophy, psychology and computer science – each of which has its own body of literature. In this section we point to a very small sample of work that has influenced our thinking.

Although a distinction may be made between ethics and morals, for our present purposes we will use the terms interchangeably. The point of an ethical theory is to provide guidelines for how to live "the good life". As such, an ethical theory should identify the nature of the good as well as provide a system for acting so as to achieve the good. Here we are not advancing a particular theory, but rather exploring how a specific cognitive architecture could support moral reasoning of different types. However, it will help readers to have at least a basic familiarity of the main ethical theories for discussions in later sections.

The three dominant ethical theories and their primary initiators include the virtue ethics of Aristotle, deontological systems descending from Kant, and consequentialist theories following from Hume, Bentham, and Mill. In virtue ethics, one lives the good life by behaving in a balanced manner, thereby exhibiting the virtues – neither too much nor too little of a given quality. Deontological systems place value on fulfilling one's duties or obligations; commonly, this is viewed as following certain rules that prescribe how one ought to behave. Consequentialist theories assign different values to different states and prescribe courses of action that result in states with the best values; utilitarianism is one particular form of consequentialism where actions are selected so as to maximize the total good across the entire population.

One approach to understanding moral reasoning is to observe and explain what people do when they are faced with morally charged decisions (Baron & Ritov, 2009; Waldmann & Dieterich, 2007). In contrast to the philosophical theories that prescriptively indicate what "ought" to be done, this approach attempts to characterize what actually takes place and why. Such studies provide hints and constraints on determining the nature of moral decision making in humans (Spranca, Minsk, & Baron, 1991). For example, many of these studies suggest that people operate using a hybrid of deontic and

consequentialist methods. The exact character of those hybrids, in terms of both structure and process, continue to be questions of interest without clear answers.

Numerous AI researchers have considered the problem of designing and implementing computational systems that embody ethical theories and perform their actions according to an ethical theory. A body of work in machine ethics has grown around this problem (M. Anderson & Anderson, 2007; Moor, 2005; Powers, 2005; Grau, 2005). Much of the interest has centered on how computational systems may be equipped to act ethically; the approaches that have been suggested naturally span the range of ethical theories. However, some have explored how the process of implementing computational models of ethical reasoning sheds light on theories of ethics in general (Guarini, 2005; Dehghani, Tomai, Forbus, & Klenk, 2008; McLaren, 2005), which is our present concern. But unlike previous efforts, which have added new mechanisms to account for moral reasoning, our work suggests that moral cognition is rather a variation of everyday cognition. In the remainder of the paper, we attempt to explain this conclusion.

## Categories of Moral Cognition

We adopt three categories of cognition related to morality. These provide a framework for thinking about computational models of morality in our ongoing and future work. These three categories may be thought of as layers, where each provides prerequisite capabilities for subsequent types of cognition. We will refer to these categories as moral *behavior*, *recognition*, and *decision making*. Our framework bears similarities to distinctions that have been made elsewhere (McNaughton, 1988; Moor, 2005; Guarini, 2005).

### Moral behavior

We start by defining *moral behavior* to be actions taken by an agent that may be evaluated by an observer as having moral value (positive or negative). That is, the action may be viewed as conforming (or not) to certain moral standards. This observer-based approach obviates the inference of motives or even cognitive states in the actor. Likewise, it potentially encompasses a broad range of behaviors as having moral values. As such, the category serves more as a foundation or starting point than as a selective and insightful distinction.

However, we stress that the category does include the universe of behaviors that we would want to consider when studying moral acts. Our observer can be either another agent in the environment in question or a privileged observer with a god's eye view. Furthermore, this definition excludes those behaviors that we would not want to consider in this context. By definition, if no observer deems a particular behavior to have moral value then we do not want to consider it in this category, although in principle any behavior can have moral valence.

### Moral interpretation

Although our concept of moral behavior need not include every conceivable action, the definition intentionally provides wide latitude in coverage. Building on this notion, we introduce *moral interpretation* as a conceptual classification, or judgment, over agent behaviors (McNaughton, 1988). Such a classification requires a cognitive capacity to recognize and distinguish moral behavior from behavior that is amoral, and in addition the ability to distinguish positive moral behavior from that which is negative. The behavior being observed and classified may be produced another agent or it may be generated by the observer itself.

Having effectively defined moral behavior as those actions over which some agent can apply moral interpretation, we should clarify the distinction between them. First, moral behavior applies to an observable action whereas moral interpretation applies to the cognitive process that makes sense of the observed action. Typically, we think of the observable action as happening in the physical environment and the cognitive process as happening in the mental states and their transitions; but technically the cognitive processes could themselves be observed and would therefore be subject to moral interpretation. Although a special case could exist in which the observed mental process is the very one making the interpretation, in general the distinction is between two quite different activities. Second, this distinction between moral interpretation and moral behavior lets us decompose the problem of designing moral agents into the problem of generating skills or behaviors that have moral associations and the problem of providing cognitive resources to appropriately interpret such behaviors. Furthermore, it underlines the conceptual component of morality apart from the behavioral component.

### Moral decision making

Once we have an agent that can recognize and categorize the behaviors exhibited by agents, we have the possibility of making choices through *moral decision making*. In this context, agents use their moral awareness during problem solving to formulate or choose intentions that reflect their own held moral values or behavioral norms. This moral reasoning would look different from agent to agent depending on the moral theories under which they are operating. For example, a deontological agent would prefer certain skills and actions while ruling out others; this framework provides search pruning at the action level. Alternatively, a consequentialist agent would adopt particular intentions based on an evaluation of the outcomes; instead of directly pruning the operators in the search space, this approach provides an evaluation function over outcomes that guides the selection of a course of action. Thus, multiple ethical theories are captured within our definition of moral reasoning.

As a consequence of this broad approach to moral cognition, we are led to consider the possibility that moral reasoning is merely a special case of everyday reasoning. We explore this question in greater depth below, and simply note here the outlines of this thinking. The mechanics of reasoning or problem solving as we think of them are unaffected by and completely unrelated to an agent's moral values or restrictions on actions. In other words, the reasoning process conducted by an agent would be identical whether it is reasoning about recharging its energy source or about helping or harming another agent. However, this equivalence of process cuts in both directions; we can also conclude from this that reasoning about how to obtain an energy recharge is actually a moral problem.

Before closing our introduction of these three categories of moral cognition, we note that our definitions primarily refer to actions – performing them, interpreting them, and planning them. We can relax this reference so as to apply our definitions to states as well. That is, we might think of moral behavior as being in a state that may be viewed by an observer as having moral value. Likewise, moral interpretation can refer to the conceptual capacity to recognize such states and moral reasoning may involve evaluating states rather than actions with respect to moral values.[1] This broader sense of these categories lets us explore moral issues in the context of ICARUS (Langley & Choi, 2006; Stracuzzi, Li, Cleveland, & Langley, 2009), an architectural theory of cognitive structures and processes to which we now turn.

## An Overview of ICARUS

We have explored moral cognition in the context of ICARUS, a unified theory of the human cognitive architecture (Newell, 1990) that imposes constraints on memory, performance processes, and learning mechanisms. Within these constraints, the framework lets one design and implement intelligent agents that accomplish a variety of tasks within many different domains. In these respects, ICARUS is similar to other cognitive architectures such as Soar (Laird, Rosenbloom, & Newell, 1986) and ACT-R (Anderson, 1993).

Among a number of distinctions from these earlier architectures, ICARUS posits separate long-term memories for storing concepts and skills. Conceptual inference is the primary means by which ICARUS agents make sense of the world. Conceptual memory comprises a set of rules, each of which specifies a head and a body, the latter containing a set of percepts, a set of conceptual relations, and a set of Boolean tests. Nonprimitive concepts may refer to other concepts in their bodies, imposing a hierarchical organization on memory. As in languages like PROLOG, the architecture infers a belief that instantiates a rule's head whenever the body matches the current situation. However, inference in ICARUS operates in a bottom-up manner that is driven by perceptions rather than by top-down queries.

The architecture's processes for skill execution, which build on the results of conceptual inference, let it carry out complex activities in the environment. Skill memory consists of a set of skills, each with a head that specifies the skill name and arguments, along with a body that states conditions for application, a set of ordered subskills, and a set of expected effects. Nonprimitive skills refer to other skills, again placing a hierarchical structure on memory. Unlike inference, the execution process operates in a top-down manner, attempting to find paths through the skill hierarchy that let the agent make progress toward completing its current task.

Skill execution supports routine activities, but it cannot handle unfamiliar situations. For this purpose, ICARUS includes a problem-solving mechanism that operates over both the skill and conceptual memories. This involves using a variant of means-ends analysis that identifies differences between the agent's current beliefs and its desired goals, retrieves skill instances that would achieve one or more of the unsatisfied goals, and selects one from this set. If the skill instance is applicable, then ICARUS executes it in the environment, generates a revised set of beliefs, and continues problem solving if necessary. If the skill's conditions are not satisfied, the architecture generates a subproblem to achieve these conditions and calls the problem solver recursively. If the system cannot retrieve any relevant skills, it chains off a conceptual clause instead. Successful solution of a subproblem leads to continuation on the original task, with ICARUS backtracking when one of its selections does not bear fruit.[2]

In addition to offering a computational theory of cognition, ICARUS provides a programming language for developing intelligent agents. Unlike other architectures, it requires that such agents operate in an external environment, typically simulated, that ground its concepts in observable percepts and that ground its skills in executable actions. ICARUS developers have constructed agents for a variety of such environments, including traditional cognitive tasks like multi-column subtraction and the Tower of Hanoi to more complex environments that we will discuss shortly. In the next section, we report our experiences with ICARUS in a number of environments that provide opportunities for moral cognition, along with their implications for this important aspect of mental processing.

[1] The distinction between event and state is less clear than it may seem; for example, qualitative reasoning about physical systems may involve thinking about extended periods of time as single states.

[2] The architecture also includes a module for learning new skills from successful problem solving, but it is not relevant to the current discussion. Langley and Choi (2006) provide more details about ICARUS' representations and mechanisms.

## Moral Cognition in Icarus

Our first foray into moral cognition involved Twig (Horswill, 2008, 2009), a physical simulator that supports a variety of object types, including humanoid agents that carry out low-level reactive behaviors like approaching a tree or picking up a nearby doll. We developed a number of Twig scenarios, one in which some people were surrounded by multiple dolls while others had none. We created Icarus concepts recognizing rich and poor people and associated skills for redistributing the wealth. As expected, this 'Robin Hood' agent repeatedly approached a wealthy person, carried one of his dolls to a less fortunate person, and left it there.

Clearly, the Icarus agent in this scenario carries out moral behavior, at least from some viewpoints, but its activities are entirely routine and rule governed. The system executes hierarchical skills in a conditional manner to carry out complex activities, but the agent does not think about their outcome, making it an example of deontic processing. However, we can also run the architecture in a different mode, where we provide the agent with one or more problems to be solved. In this case, the problem to be resolved is that no person should have fewer than one doll when others have two or more. In this setting, the Icarus problem solver detects the unsatisfied goals, retrieves high-level skills that would achieve them, executes a subset of them in turn. This variant comes closer to the consequentialist view, although the details of skill execution remain rule guided. Thus, Icarus supports a hybrid account of moral behavior that incorporates ideas from both frameworks.

Now let us consider moral interpretation, that is, recognizing whether another agent's behavior satisfies or violates one's moral tenets. We have not tested Icarus for this ability directly, but elsewhere we have reported an extension to the architecture that lets it recognize instances of complex temporal concepts (Stracuzzi et al., 2009). More specifically, we introduced mechanisms for recording episodic traces of when beliefs become true and false, along with processes for matching temporal concepts against these traces. We demonstrated their use in recognizing instances of plays in simulated football, a domain in which behavior is highly rule governed. Although this work focused on the representation and interpretation of legal plays, we could have used the same means to recognize football behavior that was illegal.

We can adapt this approach, in a fairly direct manner, to support moral interpretation. Clearly, this scheme would reflect a deontic view, since it relies centrally on using rules to determine the moral valence of an observed behavior stream. However, in related work, we have also shown how one can adapt means-ends problem solving to explain the reasons for such behavior by chaining backward from known goals through the episodic trace, which provides a consequentialist overlay. As before, we see that this traditional distinction becomes less clear when one embeds it in a cognitive architecture that offers a variety of representations and mechanisms.

The task of moral decision making, in which the agent must choose between two or more courses of action, introduces additional complications. Within Icarus, this situation arises most naturally in the context of problem solving, when the architecture must select among different skill instances that would achieve unsatisfied goals. At first glance, this appears to embody a consequentialist view, but earlier versions of Icarus made such choices randomly, and the current implementation bases them on how many goals a skill achieves and how many of its conditions match. Thus, the problem solver takes consequences into account in generating candidate actions, but not in selecting among them. Danielescu et al. (2010) explain how this can lead to undesirable situations in a simulated driving environment.

One reasonable response involves associating numeric values with conceptual predicates, including the goals and beliefs they support, as done in recent variants on the basic Icarus architecture (Choi, 2011; Asgharbeygi, Stracuzzi, & Langley, 2006). Taking these numeric annotations into account when selecting skills during problem solving would seem closer to a consequentialist treatment of moral decision making. However, Choi and Ohlsson (2010) have explored another way to guide choice in Icarus using constraints, which specify conceptual relations that should (or should not) hold under certain conditions. They have extended the architecture to carry out limited lookahead to determine whether a course of action would violate any constraints and, if so, to avoid it. This variation has a decidedly deontic character, yet one can also imagine a modulation on this idea that places numeric weights on constraints and uses them to guide decisions. This hybrid approach would, again, incorporate aspects of both moral frameworks.

Our discussion so far has dealt entirely with moral cognition that is tied to domain predicates that denote spatial relations to objects and specific physical activities. However, some moral tenets revolve around more generic relationships. Examples include the golden rule and Kant's categorical imperative not to use other humans as means to ends. These appear to require more abstract relationships that avoid reference to domain-specific predicates, and Icarus lacks both the ability to encode them or the mechanisms to operate over them. One might argue that interpretation and decision making about such abstract morals depends on a form of metacognition (Cox, 2005) that operates over traces of the agent's mental processes, rather than over descriptions of domain events. Other kinds of higher-level moral cognition, such as not misleading another intentionally or not causing unnecessary disappointment, depend on the ability to ascribe beliefs, goals, and expectations to

other agents. This is another arena in which we must extend ICARUS before it can provide complete coverage of moral cognition.

## Discussion

Experience with constructing ICARUS agents in a number of simulated environments suggests that the architecture can support routine moral behavior and at least limited forms of moral interpretation and decision making, although it also revealed that the framework cannot currently handle more abstract kinds of moral cognition. Whether other cognitive architectures like ACT-R and Soar have similar capabilities remains to be seen, but we suspect they have analogous strengths and weaknesses. Our analysis also suggested that ICARUS can model important aspects of both deontic and consequentialist views of moral cognition, although it most naturally embodies a hybrid approach that incorporates ideas from both traditions.

Our examples relied primarily on structures and processes that ICARUS already supports for other purposes. This suggests that, overall, moral cognition requires little or no additional representations or mechanisms that do not already serve another architectural need. We noted the benefits of associating numeric values with conceptual predicates in accounting for the direction and strength of moral responses, as well as uses of conditional constraints in judging the acceptability of environmental states. But again, these have been introduced into versions of ICARUS for reasons unrelated to moral cognition.

These observations are consistent with our initial claim that, despite the special treatment it has received in the literature, moral cognition does not differ in substantive ways from everyday cognition. Many will find this conclusion surprising, but we believe that it merits further consideration. An alternative phrasing that may be even more controversial is that all reasoning is an instance of moral reasoning. We hope to explore this and related issues in future research.

## Concluding Remarks

We plan to extend our models of moral cognition in several directions. We are developing an improved inference module that uses abductive reasoning to support plan understanding from partial observations, which should improve ICARUS' ability to carry out moral interpretation. We are also developing mechanisms that generate new problems in appropriate environmental situations, that prioritize them dynamically, and that abandon intentions when circumstances change or when repeated attempts have not succeeded. Most important, we plan to incorporate both numeric annotations on conceptual predicates and conditional constraints, both of which will be useful in evaluating states.

We intend to use these extensions to construct new ICARUS agents that exhibit moral interpretation and decision making in TWIG and other domains. Experience with these agents should shed additional light on the hypothesis that moral cognition and everyday cognition are, to all intents and purposes, equivalent. Nevertheless, this remains a hypothesis, and we will look for scenarios that raise genuine distinctions between the two modes of thought.

Although considerable work remains, our approach has already produced some encouraging insights. We identified three categories of moral cognition and found that each of them is supported, to some extent, by the ICARUS architecture, although we also identified aspects that it does not currently address. We also found that one can construct ICARUS agents that reflect a deontic or consequentialist view, as well as hybrid models that incorporate ideas from both frameworks. We hope that future research will let us clarify and expand on these initial insights.

## Acknowledgements

This material is based on research sponsored by ONR under agreement N00014-10-1-0487. We thank Paul Bello and Mark Nelson for discussions that improved the content of the paper, and we thank the anonymous reviewers for their comments. The views and conclusions contained herein are the authors' and should not be interpreted as representing the official policies or endorsements of ONR or the U. S. Government, either expressed or implied.

## References

Anderson, J. (1993). *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum.

Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine*, *28*(4), 15–26.

Asgharbeygi, N., Stracuzzi, D., & Langley, P. (2006). Relational temporal difference learning. In *Proceedings of the Twenty-Third International Conference on Machine Learning* (pp. 49–56). Pittsburgh, PA: IMLS.

Baron, J., & Ritov, I. (2009). Protected values and omission bias as deontological judgments. In D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin (Eds.), *Moral judgment and decision making.* San Diego, CA: Academic Press.

Choi, D. (2011). Reactive goal management in a cognitive architecture. *Cognitive Systems Research*, *12*, 293–308.

Choi, D., & Ohlsson, S. (2010). Learning from failures for cognitive flexibility. In *Proceedings of the Thirty-Second Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Cox, M. T. (2005). Metacognition in computation: A selected history. In *Proceedings of the AAAI Spring Symposium on Metacognition in Computation* (pp. 1–17). Stanford, CA: AAAI Press.

Danielescu, A., Stracuzzi, D. J., Li, N., & Langley, P. (2010). Learning from errors by counterfactual reasoning in a unified cognitive architecture. In *Proceedings of the Thirty-Second Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Dehghani, M., Tomai, E., Forbus, K., & Klenk, M. (2008). An integrated reasoning approach to moral decision-making. In *Proceedings of the Twenty-Third Conference on Artificial Intelligence.* Menlo Park, CA: AAAI Press.

Grau, C. (2005). There is no "I" in "robot": Robotic utilitarians and utilitarian robots. In *Machine Ethics: Papers from the 2005 AAAI Fall Symposium.* Menlo Park, CA: AAAI Press.

Guarini, M. (2005). Particularism and generalism: How AI can help us to better understand moral cognition. In *Machine Ethics: Papers from the 2005 AAAI Fall Symposium.* Menlo Park, CA: AAAI Press.

Horswill, I. (2008). Lightweight procedural animation with believable physical interactions. In *Proceedings of the Fourth Artificial Intelligence and Interactive Digital Entertainment Conference.* AAAI Press.

Horswill, I. (2009). Very fast action selection for parameterized behaviors. In *Proceedings of the Fifth International Conference on Foundations of Digital Games.* Orlando, FL: ACM Press.

Laird, J. E., Rosenbloom, P. S., & Newell, A. (1986). Chunking in Soar: The anatomy of a general learning mechanism. *Machine Learning*, *1*, 11–46.

Langley, P., & Choi, D. (2006). Learning recursive control programs for problem solving. *Journal of Machine Learning Research*, *7*, 493–518.

McLaren, B. M. (2005). Lessons in machine ethics from the perspective of two computational models of ethical reasoning. In *Machine Ethics: Papers from the 2005 AAAI Fall Symposium.* Menlo Park, CA: AAAI Press.

McNaughton, D. (1988). *Moral vision: An introduction to ethics.* Oxford, UK: Wiley-Blackwell.

Moor, J. H. (2005). The nature and importance of machine ethics. In *Machine Ethics: Papers from the 2005 AAAI Fall Symposium.* Menlo Park, CA: AAAI Press.

Newell, A. (1990). *Unified theories of cognition.* Cambridge, MA: Harvard University Press.

Powers, T. M. (2005). Deontological machine ethics. In *Machine Ethics: Papers from the 2005 AAAI Fall Symposium.* Menlo Park, CA: AAAI Press.

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology*, *27*, 76–105.

Stracuzzi, D. J., Li, N., Cleveland, G., & Langley, P. (2009). Representing and reasoning over time in a cognitive architecture. In *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Waldmann, M. R., & Dieterich, J. (2007). Throwing a bomb on a person versus throwing a person on a bomb: Intervention myopia in moral intuitions. *Psychological Science*, *18*(3), 247–253.