An Integrated Reasoning Approach to Moral Decision-Making

Morteza Dehghani, Emmett Tomai, Ken Forbus, Matthew Klenk

Qualitative Reasoning Group, Northwestern University
2133 Sheridan Road, Evanston, IL 60208 USA
Contact: morteza@northwestern.edu

Abstract

We present a computational model, MoralDM, which integrates several AI techniques in order to model recent psychological findings on moral decision-making. Current theories of moral decision-making extend beyond pure utilitarian models by relying on contextual factors that vary with culture. MoralDM uses a natural language system to produce formal representations from psychological stimuli, to reduce tailorability. The impacts of secular versus sacred values are modeled via qualitative reasoning, using an order of magnitude representation. MoralDM uses a combination of first-principles reasoning and analogical reasoning to determine consequences and utilities when making moral judgments. We describe how MoralDM works and show that it can model psychological results and improve its performance via accumulating examples.

Introduction

While traditional models of decision-making in AI have focused on utilitarian theories, there is considerable psychological evidence that these theories fail to capture the full spectrum of human decision-making. In particular, research on moral reasoning has uncovered a conflict between normative outcomes and intuitive judgments. This has led some researchers to propose the existence of deontological moral rules; i.e., some actions are immoral regardless of consequences, which could block utilitarian motives. Consider the starvation scenario (from Ritov & Baron 1999) below:

> A convoy of food trucks is on its way to a refugee camp during a famine in Africa. (Airplanes cannot be used.) You find that a second camp has even more refugees. If you tell the convoy to go to the second camp instead of the first, you will save 1000 people from death, but 100 people in the first camp will die as a result.
>
> Would you send the convoy to the second camp?

The utilitarian decision would send the convoy to the second camp, but participants tended to not divert the truck.

Making these types of decisions automatically requires an integrated approach, including natural language understanding, qualitative reasoning, analogical reasoning, and first-principles reasoning. This paper describes a cognitively motivated model of moral decision-making, called MoralDM, which incorporates two modes of decision-making: utilitarian and deontological. To reduce tailorability, a natural language understanding system is used to semi-automatically produce formal representations from psychological stimuli re-rendered in simplified English. The different impacts of secular versus sacred values are modeled via qualitative reasoning, using an order of magnitude representation. MoralDM combines first-principles reasoning and analogical reasoning to implement rules of moral decision-making and utilize previously made decisions. We evaluate our system against established psychological results.

We begin by summarizing relevant psychological results and background. Next, we describe MoralDM and how it works. Then we show that MoralDM can account for results from two psychological studies. An analysis of its performance demonstrates learning through accumulation of examples and the importance of integrated reasoning. Finally, we discuss related and future work.

Moral Decision-Making

Morality as a topic of experimental scientific inquiry has attracted the attention of psychologists for more than eight decades. After the initial domination of rational approaches to decision-making, the conflict between normative outcomes and intuitive judgments led some researchers to suggest the existence of *sacred values*, which are not allowed to be traded off, no matter the consequences (Baron and Spranca, 1997). These sacred values are known to block utilitarian motives by evoking deontological moral rules. In our example, given that life is a sacred value, people often refuse to take an action which would result in taking lives. Tetlock (2000) defines sacred values as “those values that a moral community treats as possessing transcendental significance that precludes comparisons, trade-offs, or indeed any mingling with secular values”. People who have sacred or protected values tend to reject the need for tradeoffs, no matter the consequences, and often show strong emotional reactions, such as anger, when these values are challenged.

When sacred values are involved, people tend to be concerned with the nature of their action rather than the utility of the outcome. Baron and Spranca (1997) argue that people show lower *quantity sensitivity* to outcome utilities when dealing with sacred values. That is, they become less sensitive to the consequences of their choices, leading them to prefer inaction, even if it results in a lower outcome utility, over an action which violates a sacred value. The degree of outcome sensitivity varies with culture and the context of the scenario. Lim and Baron (1997) show that people in different cultures tend to protect different values and demonstrate different levels of sensitivity towards shared sacred values.

In addition to sacred values, the causal structure of the scenario affects people's decision-making. Waldmann and Dieterich (2007) show that people act more utilitarian, i.e., become more sensitive to action outcome utilities, if their action influences the agent of harm rather than the potential patient.

MoralDM

Moral decision-making is a complex reasoning process. In psychological studies, scenarios are presented to human subjects in natural language. The research summarized above identifies a number of contextual factors which cause subjects to become less sensitive to the outcome utilities of their decisions. Other research has also shown that analogy plays a role in many decision-making domains. Consequently, a model of moral decision-making needs to include natural language understanding, a method for comparing outcome quantities that takes into account quantity sensitivity, a method for reasoning about outcomes utilities and sacred values, and the ability to utilize previous decisions or examples when reasoning about new situations.

Our model of moral decision-making, MoralDM, incorporates two mutually exclusive modes: utilitarian and deontological. If there are no sacred values involved in the case being analyzed, MoralDM applies traditional rules of utilitarian decision-making by choosing the action which provides the highest outcome utility. On the other hand, if MoralDM determines that there are sacred values involved, it operates in deontological mode and becomes less sensitive to the outcome utility of actions, preferring inactions to actions.

MoralDM has been implemented using the FIRE reasoning engine. The knowledge base contents are a 1.2 million fact subset of Cycorp's ResearchCyc[1] knowledge base, which provides formal representations about everyday objects, people, events and relationships. The KB also includes representations we have developed to support qualitative and analogical reasoning. The KB provides a formal ontology useful for representing and reasoning about moral decision-making scenarios.

Figure 1 provides an overview of the MoralDM architecture. To solve a given moral decision-making scenario, MoralDM begins by using EA NLU, a natural language understanding system, to semi-automatically translate simplified English scenarios into predicate calculus. Given this representation, the presence of sacred values and relevant contextual factors are computed via a fixed set of rules. A number of known sacred values are stored in the KB. For a new scenario a set of rules are applied to decide whether the case includes sacred values or not. The orders of magnitude reasoning module (OMR) then calculates the relationship between the utility of each choice. Using the outcome of the orders of magnitude reasoning module, MoralDM utilizes a hybrid reasoning approach consisting of a first-principles reasoning module (FPR) and an analogical reasoning module (AR) to arrive at a decision. The first-principles reasoning module suggests decisions based on rules of moral reasoning. The analogical reasoning module compares a given scenario with previously solved decision cases to determine whether sacred values exist in the new case and suggest a course of action. We believe using hybrid reasoning both gives the system the ability to tackle a boarder range of decision-making scenarios and provides a more cognitively plausible approach to decision-making.

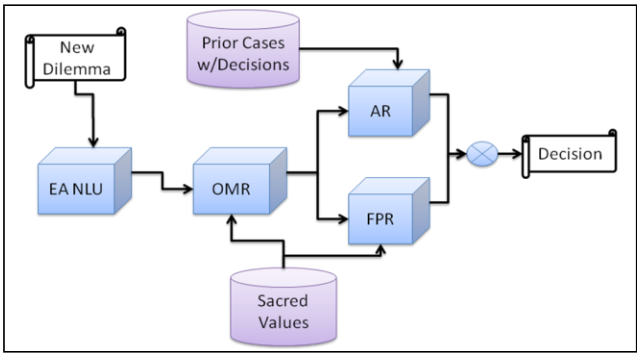

Figure 1: MoralDM Architecture

The first-principles and analogical reasoning modules work in parallel and complement each other by providing support (or disagreement) for a decision. If both succeed and agree, the decision is presented. When one module fails to arrive at a decision, the answer from the other module is used. If the modules do not agree, the system selects the first-principles reasoning module's choice. If both fail, the system is incapable of making a decision. After a decision is made for a given scenario, it can be stored in the case library for future use. This enables the system to make decisions in more scenarios as it accumulates experience. Next, we discuss each module in detail.

[1] Research.cyc.com

Explanation Agent NLU

Our inputs are dilemmas from the psychological literature, expressed in natural language. To construct formal representations of these stimuli, we have extended the Explanation Agent Natural Language Understanding system (EA NLU, Kuehne and Forbus, 2004). Because unrestricted automatic natural language understanding is beyond the state of the art, EA NLU uses simplified language and operates semi-automatically. This practical approach allows us to broadly handle syntactic and semantic ambiguities and to build deep representations suitable for complex reasoning. This is a significant advantage over having experimenters construct representations entirely by hand for two reasons. First, constructing representations by hand is very time-consuming and requires substantial expertise. Second, hand-coding increases tailorability, i.e., the possibility that representation choices were made to get a particular example to work, as opposed to being uniform, independently motivated conventions. Since EA NLU is used by multiple projects and relies on an off-the-shelf knowledge base, tailorability is greatly reduced.

EA NLU uses Allen's bottom-up chart parser (Allen, 1995) in combination with the COMLEX lexicon (Macleod *et al.* 1998) and a simplified English grammar (Kuehne and Forbus, 2004). The parser uses subcategorization frames from ResearchCyc for word and common phrase semantics. Compositional frame-based semantics from the parsing process are transformed using dynamic logic principles from Discourse Representation Theory (DRT) (Kamp and Reyle, 1993). The resulting set of *discourse representation structures* (DRS) supports numerical and qualitative quantification, negation, implication, modal embedding and explicit and implicit utterance sub-sentences.

```
A convoy of trucks is transporting food
to a refugee camp during a famine in
Africa. 1000 people in a second refugee
camp will die. You can save them by
ordering the convoy to go to that
refugee camp. The order will cause 100
people to die in the first refugee camp.
```

Figure 2: Starvation scenario in simplified English

Figure 2 contains the simplified English for the starvation scenario. Given these statements, EA NLU identifies events (e.g. dying, ordering, saving), entities (e.g. two quantified sets of people, the convoy) and an explicit reference to the listener, "you". Figure 3 contains the frame-based interpretation of the order. This set of facts is contained within a DRS which is modally embedded with the operator possible.

Causal links are explicitly stated between the order and the saving and the order and the second set of deaths. The abstraction of saving drives inferential attention to events in the description that the beneficiary may be being saved from. The expected future modality of the first set of deaths makes it a reasonable candidate.

```
(isa order131049 Ordering-CommunicationAct)
(performedBy order131049 you128898)
(recipientOfInfo order131049 convoy127246)

(infoTransferred order131049
 (and
  (isa refugee-camp129739 RefugeeCamp)
  (isa convoy127246 Convoy)
  (isa go129115 Movement-TranslationEvent)
  (primaryObjectMoving go129115 convoy127246)
  (toLocation go129115 refugee-camp129739)))
```

Figure 3: Predicate calculus for ordering

Based on the possible modality of the saving/ordering sequence, combined with the use of the explicit reference to the listener, the system infers an abstraction of choice being presented with known consequences resulting from both action and inaction. Figure 4 contains the inferred abstraction of choice and its causal consequences.

```
(isa Sel131949 SelectingSomething)
(choices Sel131949 order131049)
(choices Sel131949 Inaction131950)
(causes-PropSit
  (chosenItem Sel131949 Inaction131950)
  die128829)
(causes-PropSit
  (chosenItem Sel131949 order131049)
  save128937)
```

Figure 4: Predicate calculus for the choice presented

Order of Magnitude Reasoning Module

We model people's degree of quantity sensitivity toward outcomes by using Dague's (1993) ROM(R) relative order of magnitude formalism. Order of magnitude representations provide the kind of stratification that seems necessary for modeling the impact of sacred values on reasoning. One of the features of ROM(R) is that it includes a parameter, *k,* which can be varied to capture differences in quantity sensitivity. We implemented a simplified version of ROM(R) using one degree of freedom, *k*, resulting in three binary relations which can be computed using the following rules:

- $A \approx_k B \Leftrightarrow |A\text{-}B| \leq k * \mathrm{Max}(|A|,|B|)$
- $A <_k B \Leftrightarrow |A| \leq k * |B|$
- $A \neq_k B \Leftrightarrow |A\text{-}B| > k * \mathrm{Max}(|A|,|B|)$

These relations respectively map to close to, greater than and distant from. *k* can take any value between 0 and 1, with a higher *k* resulting in less quantity sensitivity. Depending on the sacred values involved and the causal structure of the scenario, we vary *k* to capture sensitivity towards the utility of the outcome.

The inputs to the orders of magnitude reasoning module include the sacred values for the culture being modeled and the causal structure of the scenario. Using the predicate calculus produced by EA NLU, the module calculates the expected utility of each choice by summing the utility of its consequences. For each consequence of a choice, it uses its rules to ascertain if the outcome is a positive or negative outcome, and to identify any sets whose cardinality matters in the decision (e.g., number of people at risk).

After computing utilities, the orders of magnitude reasoning module selects a k value based upon the context of the scenario. Assuming that the relationship between the utilities, a and b, are comparable, MoralDM sets k to $1 - (|a| / |b|)$. This results in the relationship between the utilities falling within <, right between ≠ and ≈. We model the impact of sacred values via a modifier ε: when $k < ½$, ε is $k/(1 - k)$, and, when $k \geq ½$, ε is $(1 - k)/ k$. If the decision involves a sacred value for the modeled culture, setting k to $k + \varepsilon$ shifts the relationship between utilities from greater than to close to, resulting in the system being less sensitive to the numeric utility of the outcome. On the other hand, if there are no sacred values involved, the system substitutes k with $k - \varepsilon$ thereby making the system more quantity sensitive to the computed utilities. In addition to sacred values, the causal structure of the scenario affects k. The orders of magnitude reasoning module checks to see if the scenario contains patient intervention or agent intervention. It uses low quantity sensitivity for agent intervention and high otherwise, consistent with psychological findings (Waldmann and Dieterich 2007). The system also checks for direct versus indirect causation. In the case of indirect causation, a higher degree of sensitivity is applied. Returning to the starvation scenario, there are two choices: ordering and inaction. For ordering, there are two consequences, 1000 people in the second camp will be saved and 100 people in the first camp will die. Consulting the KB, the system determines that dying has negative utility and saving positive, resulting in a choice utility of 900 for the ordering choice. Using the same procedure, the utility for inaction is calculated to be -900. Using the formula given above, k is initially set to 0 with $\varepsilon = 1$. Given that both choices involve agent intervention and indirect causation, there are no structural differences between the two choices. Therefore, the k value is set solely by the existence of sacred values. In this case, causing someone to die is a sacred value resulting in k being set to $k + \varepsilon = 1$, therefore causing the system to act less quantity sensitive. Using ROM(R), the relationship between the utilities of the two choices is calculated to be ≈. On the other hand, if there had not been a sacred value, the value of k would have remained 0 causing the relationship between the utilities to be ≠. These utilities, 900 and -900, and the computed relationship, ≈, are provided to the first-principles and analogical reasoning modules.

First-Principles Reasoning Module

Motivated by moral decision-making research, the first-principles reasoning module makes decisions based upon the orders of magnitude relationship between utilities, sacred values, computed utilities, and action vs. inaction. It uses three methods for making decisions. First, the utilitarian method, which selects the choice with the highest utility, is invoked when the choice does not involve a sacred value. Second, in situations with sacred values and without an order of magnitude difference between outcomes, the pure-deontological method selects the choice that does not violate a sacred value. Third, when the scenario contains sacred values and an order of magnitude difference between outcomes, the utilitarian-deontological method selects the choice with the highest utility. Therefore, the pure-deontological method is the only method that makes decisions that violate utilitarian norms.

In the starvation scenario, there is a sacred value, people dying, and no order magnitude difference between the utility of the two choices. Therefore, the system uses the pure deontological method and selects the inaction choice.

These methods are mutually exclusive, returning at most one choice per scenario. Given the breadth of moral reasoning scenarios, the rules implementing the first-principles reasoning module are not complete. Therefore, it necessarily fails on some scenarios. These cases highlight the need for the hybrid-reasoning approach taken in MoralDM. The resulting choice is compared with the results of the analogical reasoning module of MoralDM.

Analogical Reasoning Module

Analogy plays important roles in decision-making. When making a choice, decision makers frequently use past experiences and draw inferences from their previous choices (Markman and Medin, 2002). Research on the use of analogy in decision-making suggests that the comparison between a target and a base involves an alignment process, where structural relations are weighted more heavily than surface similarities (Gentner and Markman, 1997).

To model analogy in decision-making, we use the Structure-Mapping Engine (SME) (Falkenhainer *et al.* 1989), a computational model of similarity and analogy based on Gentner's (1983) structure mapping theory of analogy in humans. SME operates over structured representations, consisting of entities, attributes of entities and relations. There are both first-order relations between entities and higher-order relations between statements. Given two descriptions, a *base case* and a *target case*, SME aligns their common structure to find a mapping between the cases. This mapping consists of a set of correspondences between entities and expressions in the two cases. SME produces mappings that maximize *systematicity*; i.e., it prefers mappings with higher-order relations and nested relational structure. The *structural evaluation score* of a mapping is a numerical measure of similarity between the base and target. SME identifies elements in the base that fail to map to the target and uses the common relational structure to calculate *candidate inferences* by filling in missing structures in target.

Running concurrently with the first-principles reasoning module, the analogical reasoning module uses comparisons between new cases and previously solved cases to suggest decisions. When faced with a moral decision scenario, the analogical reasoning module uses SME to compare the new case with every previously solved scenario in its memory. The similarity score between the novel case and each solved scenario is calculated using SME by normalizing the structural evaluation score against the size of the scenario. If this score is higher than a certain threshold and both scenarios contain the same order of magnitude relationship between outcome utilities, then the candidate inferences are considered as valid analogical decisions. If the scenarios have different orders of magnitude relationships, it is likely that a different mode of reasoning should be used for the target scenario and the analogical reasoning module rejects the analogical inference. After comparing against all of the solved scenarios, the analogical reasoning module selects the choice in the new scenario supported by the largest number of analogs. In the case of a tie, the analogical reasoning module selects the choice with the highest average similarity score supporting it. Because analogical alignment is based upon similarities in structure, similar causal structures and/or sacred values align similar decisions. Therefore, the more structurally similar the scenarios are, the more likely the analogical decision is going to be the correct moral one. If there are no previously solved cases in the case library, or they fail the filtering criterion above, the analogical reasoning module fails to come up with an answer. Therefore, the first-principles reasoning module is needed to bootstrap the analogical reasoning module.

Returning to our starvation example, the analogical reasoning module can solve this decision problem through an analogy with a traffic scenario given below, in which the system chose to not transfer funds:

> A program to combat accidents saves 50 lives per year in a specific area. The same funds could be used to save 200 lives in another area, but the 50 lives in the first area would be lost.
>
> Do you transfer the funds?

The analogical decision is determined by the candidate inferences where the decision in the base, inaction, is mapped to the choice representing inaction in the target. Because the traffic scenario contains the same the order of magnitude relationship, ≈, as in the starvation scenario, the system accepts the analogical decision.

Integrated System Evaluation

We conducted a series of experiments to evaluate MoralDM. Experiments 1 and 2 evaluate MoralDM as a model for moral decision-making and illustrate the importance of using both analogical and first-principles reasoning. In these two experiments, there are cases where one of the reasoning modules fails, but MoralDM is still able to give the correct decision by using the other module. Experiment 3 investigates the claim that additional decision cases improve the analogical reasoning module.

Experiment 1

We evaluated MoralDM by running it on 8 moral decision-making scenarios taken from two psychology studies (Waldmann and Dieterich, 2007; Ritov and Baron, 1999). In all the scenarios used, traditional utility theories fail to predict subjects' responses, because subjects select the choice which provides a smaller overall outcome utility. We compare MoralDM's decisions to subjects' responses in these experiments. If the decision of MoralDM matched those of the subjects, as reported by the authors, we consider it a correct choice.

For each case, EA NLU semi-automatically translated the simplified English version of the original psychology scenario into predicate calculus. The sacred values and the relevant contextual factors are computed via rules. Then the order of magnitude reasoning module calculated the relative relation between the utilities. This relation and the sacred values involved in the case were sent to the first-principles and analogical reasoning modules. Correct decisions are then added to MoralDM's experiences.

The analogical reasoning module failed to choose the correct decision in three cases. As discussed previously, this module failed on the first case because it did not have any cases in its memory to reason from. The other two cases involved scenarios for which no appropriate analog could be found due to their considerably different causal structure. In all three cases, the first-principles module made the correct decision. Overall, MoralDM made the correct choice in all of the scenarios ($p < 0.005$).

Experiment 2

One of the more difficult aspects in building the first-principles reasoning module is the number of rules required to handle the broad range of situations covered in moral decision-making. This experiment is designed to test the hypothesis that the analogical reasoning module is capable of making moral decisions in situations when gaps in the knowledge base or rule set prevent the first-principles reasoning module from making a decision. In this experiment, all 12 moral decision-making scenarios from Ritov and Baron (1999) were used as inputs. Unlike the other experiments, 8 could not be translated by EA NLU, so we encoded those manually.

	# of correct decisions
MoralDM	11 (92%)
First-principles	8 (67%)
Analogical Reasoning	11 (92%)

Table 1: MoralDM results

Table 1 displays the results of MoralDM, broken down by reasoning modules. In 8 scenarios, both modules provide the correct answer. In three scenarios, the first-principles reasoning module fails to make a prediction, but the analogical reasoning module provides the correct answer. In one scenario, both modules fail.

The first-principles reasoning module fails in four of the scenarios because our current rules for handling cases with unique structure or content are limited. For example, there is a scenario about Israeli settlements where the first-principles module fails. The system does not have the necessary rules to determine that Israeli land is considered a sacred value for Israelis, and it cannot be traded off. However, the analogical reasoning module is still able to make decisions in three of these cases based upon similarities with other scenarios, e.g. with a scenario where saving a nature preserve was a sacred value. The analogical reasoning module fails on the fourth case because the causal structure of the case is very different from the other cases.

Experiment 3

This experiment addresses the question of how effective the analogical reasoning module is at learning from experience. We measure how performance is affected as a function of the number of previously solved cases in memory. Given the 8 solved scenarios from experiment 1, we created case libraries of every combination of these scenarios. This provided us with 254 different case libraries (8 of size 1, 28 of size 2, 56 of size 3…). Then, with each case library, we tested the analogical reasoning module by running it on each of the scenarios not in the case library. So for each of the 8 libraries of size 1, the test consisted of 7 decision scenarios for a total of 56 decision scenarios.

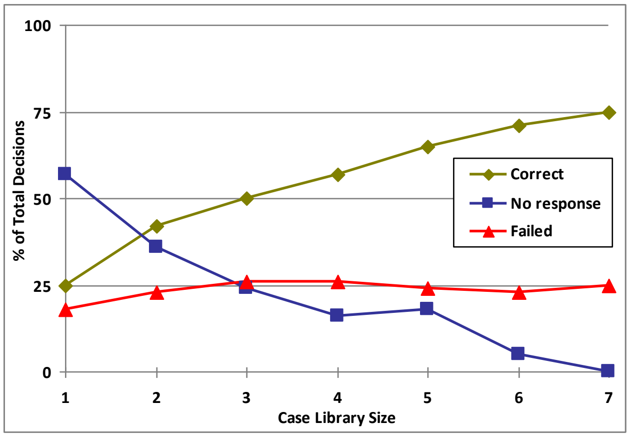

Figure 5: Analogical reasoning results

Figure 5 shows the performance of the analogical reasoning module as a function of the number of available cases. There is a monotonic increase in the number of correct answers as the size of the library increases (Pearson’s r = .97, p < .0001). Also, there is a significant decrease in the number of cases where the analogical reasoning module does not come up with an answer (r = -.95, p < .001). The number of incorrect decisions changes insignificantly from 18% to 25% (r = .53, p < .22). The statistics reported have been computed by comparing each series against the size of the case library.

Discussion

The results of these experiments are very encouraging. As shown in experiments 1 and 2, our system matches human behavior on a set of decision-making scenarios. This result would not be possible without the integrated approach. First, the input was given in natural language requiring EA NLU. Second, these cases all involved sacred values; therefore the orders of magnitude reasoning module’s computed relationship between outcome utilities is essential to providing the correct answer. Third, the first-principles and analogical reasoning modules were both needed to select the appropriate action.

We believe that the breadth of moral decision-making means that the integration of analogical and first principles reasoning has significant benefits. Without both analogical and first principles reasoning, MoralDM would have failed on a considerable number of problems from the first two experiments. In experiment 1, we demonstrated the necessity of the first-principles reasoning module where there are insufficient appropriate prior cases for analogical reasoning. The analogical reasoning module alone could not have correctly answered the 8 cases. In experiment 2, we demonstrated that the analogical reasoning module enables the system to handle a wider range of decision-making scenarios where gaps in the knowledge base and/or rule set prevent the first-principles reasoning module from answering correctly. Without the analogical reasoning module, MoralDM would have failed on three more cases.

Experiment 3 provides additional support for the importance of the analogical reasoning module within MoralDM and demonstrates how it serves as a learning component. The results show a significant improvement in the analogical reasoning module’s performance as the number of cases in MoralDM’s memory increased.

Related Work

Several research projects have focused on building ethical advisors. The MedEthEx system uses ILP techniques to learn decision principles from training cases (Anderson *et al.* 2006). Mclaren's Truth-Teller and SIROCCO systems (2005) use case-based reasoning to highlight relevant ethical considerations and arguments to a human user. Like them, we use prior cases, but to guide the system’s own reasoning, rather than give advice. They also were not designed to model the effects of sacred versus secular values that MoralDM captures.

Computational models of cultural reasoning are receiving increasing attention. For example, the CARA system (Subrahmanian *et al.* 2007) is part of a project to “understand how different cultural groups today make decisions and what factors those decisions are based upon”. CARA uses semantic web technologies and opinion extraction from weblogs to build cultural decision models consisting of qualitative rules and utility evaluation. While we agree that qualitative reasoning must be integrated with traditional utility evaluation, we also believe that analogy plays a key role in moral reasoning. Moreover, we differ by evaluating MoralDM against psychological studies, which helps ensure its judgments will be like those that people make.

Our combination of analogical and first-principles reasoning is inspired in part by Winston's (1982) use of both precedents and rules to reason about a situation. His work was hampered by the lack of off-the-shelf large-scale knowledge bases, and the technologies for NLU and analogical reasoning have improved since then.

Conclusions and Future Work

MoralDM integrates multiple AI techniques to model human moral decision-making. It uses qualitative modeling to reason about utilities, capturing differences between sacred and secular values via an order of magnitude representation. It uses a combination of first-principles logical reasoning and analogical reasoning to determine the utility of outcomes and make decisions based on this information, producing answers in a wider range of circumstances than either alone can handle. Natural language input of scenarios, in simplified English, reduces tailorability, a key problem in cognitive simulation research. In a series of experiments we showed the necessity of integrating all the above modules. While there is still more to be done, we think MoralDM represents an important step in computational modeling of moral decision-making.

We plan to pursue several lines of investigation next. First, we plan to extend the valuation rules to model different cultures, based on existing collaborations with cognitive psychologists and anthropologists. Second, we plan to extend the range of EA NLU coverage to handle a wide range of cultural stories. This will enable us to create story libraries for different cultural groups, and translate transcripts from interview data more easily. Third, we plan to incorporate a cognitively plausible model of similarity-based retrieval, MAC/FAC (Forbus *et al*. 1995), to make analogical reasoning more scalable as the story library grows. Finally, we plan to test MoralDM on a wider range of problems, using data gathered from participants from multiple cultural groups.

Acknowledgments

This research was supported by an AFSOR MURI. We thank Andrew Lovett for many useful insights.

References

Allen, J. F. 1995.. *Natural Language Understanding.* (2nd ed). Redwood City, CA.: Benjamin/Cummings.

Anderson, M., Anderson, S., and Armen, C. 2006. An Approach to Computing Ethics. *IEEE Intelligent Systems*. 21(4): 56-63

Baron, J., and Spranca, M. 1997. Protected Values. *Organizational Behavior and Human Decision Processes* 70: 1–16.

Dauge, P. 1993. Numeric Reasoning with Relative Orders of Magnitude. In *Proceedings of QR93.*

Falkenhainer, B., Forbus, K. and Gentner, D. 1989. The Structure-Mapping Engine: Algorithms and Examples. *Artificial Intelligence,* 41: 1-63.

Forbus, K., Gentner, D. and Law, K. 1995. MAC/FAC: A Model of Similarity-based Retrieval. *Cognitive Science*, 19(2), 141-205.

Gentner, D. 1983. Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Science*, 7: 155-170.

Gentner, D., & Markman, A. B. (1997). Structural alignment in analogy and similarity. American Psychologist, 52(1), 45-56.

Kamp, H. and Reyle, U. 1993. *From Discourse to Logic*, Kluwer, Dordrecht.

Kuehne, S. and Forbus, K. 2004. Capturing QP-Relevant Information from Natural Language Text. *Proceedings of QR04.*

Lim, C. S, and Baron, J. (1997), Protected values in Malaysia Singapore, and the United States. *Manuscript, Department of Psychology*, University of Pennyslvania.

Markman, A and Medin, D.L. 2002. *Decision Making.* Stevens Handbook of Experimental Psychology, 3rd edition: Volume 2, Memory and Cognitive Processes. New York: Wiley.

Macleod, C., Grishman, R., and Meyers, A. 1998. COMLEX Syntax Reference Manual, Version 3.0. Linguistic Data Consortium. University of Pennsylvania: Philadelphia, PA.

McLaren, B. 2005. Lessons in Machine Ethics from the Perspective of Two Computational Models. In Anderson, M. *et al.* (Eds.) *Machine ethics: Papers from the AAAI Fall Symposium*, Technical Report FS-05-06. Menlo Park, CA.

Ritov, I. and Baron, J. 1999. Protected Values and Omission Bias. *Organizational Behavior and Human Decision Processes*, 79(2): 79-94.

Subrahmanian, VS., Albanese, M., Martinez, V., Nau, D., Reforgiato, D., Simari, G., Sliva, A., and WIlkenfeld, J. 2007. CARA: A Cultural Adversarial Reasoning Architecture. *IEEE Intelligent Systems.* 22(2): 12-16.

Tetlock, P. E. (2000). Cognitive biases and organizational correctives: Do both disease and cure depend on the ideological beholder? A*dministrative Science Quarterly*, 45(2), 293-326.

Waldmann, M. R., and 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.

Winston, P.H. 1982. Learning New Principles from Precedents and Exercises. *Artificial Intelligence* 19(3), 321-350.