Moral Decision-Making by Analogy: Generalizations vs. Exemplars

Joseph A. Blass, Kenneth D. Forbus
Qualitative Reasoning Group, Northwestern University
2133 Sheridan Road, Evanston, IL 60208 USA
Contact: joeblass@u.northwestern.edu

Abstract

Moral reasoning is important to model accurately as AI systems become ever more integrated into our lives. Moral reasoning is rapid and unconscious; analogical reasoning, which can be unconscious, is a promising approach to model moral reasoning. This paper explores the use of analogical generalizations to improve moral reasoning. Analogical reasoning has already been used successfully to model moral reasoning in the MoralDM model, but it exhaustively matches across all known cases, which is computationally intractable and cognitively implausible for human-scale knowledge bases. We investigate the performance of an extension of MoralDM to use the MAC/FAC model of analogical retrieval over three conditions, across a set of highly confusable moral scenarios.

Introduction

Research on moral reasoning and decision-making in humans has revealed that certain moral decisions are based on moral rules rather than utilitarian considerations. These rules are concerned with the morality of actions rather than their outcomes (Baron and Spranca 1997). Furthermore, people are sensitive to structural differences, leading us to make different judgments even when considering cases that seem similar on the surface. Consider these two scenarios (from Waldmann and Dietrich 2007) below:

> Bomb 1: "In a restaurant, a bomb threatens to kill 9 guests. The bomb could be thrown onto the patio, where 1 guest would be killed [but the 9 would be spared]."

> Bomb 2: "In a restaurant, a bomb threatens to kill 9 guests. One [separate] guest could be thrown on the bomb, which would kill this one guest [but the 9 would be spared]."

Despite the two scenarios having strong surface and structural similarities as well as identical utilitarian considerations, participants in this study generally said that taking the proposed action was morally acceptable in the first scenario but not in the second scenario.

In addition to making non-utilitarian moral decisions, humans tend to be able to judge the moral value of actions and situations quickly and with little conscious reasoning; in fact, there is evidence that justifications for moral judgments come after the judgment has been made (Haidt 2001). Furthermore, research suggests that our moral rules are not innate or learned by rote, but are acquired culturally from exposure to moral narratives (Darley and Shultz 1990). These features suggest that purely first-principles, rule-based reasoning is not sufficient to account for moral reasoning.

Analogical reasoning involves comparing structured cases. Analogical reasoning can be unconscious (Day and Gentner 2003), and is therefore an intriguing approach for modeling moral decision-making. However, analogical retrieval often results in a match with strong surface similarities, rather than structural or relational similarities, although experts do not show this pattern (Novick 1988). This poses a problem for cases such as the two bomb scenarios, since each scenario is most surface-similar to the other, and therefore the likely target of a match for someone who knows one scenario and is evaluating the other.

A potential solution is to use analogical generalizations, built from similar-enough exemplars, which emphasize what is common between them while deprecating facts unique to individual exemplars. If presented in an order conducive to assimilation, generalizations will be formed on the basis of higher-order relations shared across cases (Kuehne et al. 2000); as the generalizations grow, these facts become more important than facts unique to cases. The Sequential Analogical Generalization Engine (SAGE), has

been shown to facilitate analogical reasoning (McLure, Friedman, and Forbus 2010; Chang and Forbus 2013).

This paper shows how reasoning by analogy using generalizations of solved moral scenarios, rather than across those individual cases, leads to better performance as the size of the training case-library grows. Forming these generalizations is important as it may better represent the way we learn and reason about moral dilemmas. It will eventually be important for autonomous moral reasoning systems to *learn* moral rules, not just be told them. Firm rules can be brittle and contradict each other (Iliev et al. 2009), yet humans do not stop making consistent moral decisions; our systems need to learn and reason the same way we do in moral domains if we are ever to trust them to make the right decision in high-stakes environments, such as driving a car on the highway.

This work is a continuation of research done on MoralDM, a model of moral judgment (Dehghani et al. 2008a, Dehghani et al. 2008b). We begin by reviewing the relevant psychological background and findings on moral decision-making and analogical reasoning. Next we briefly describe previous work on MoralDM. Then we show how an analogy-only adaptation of that system performs using four analogy-based conditions across different sizes of case libraries. Finally, we discuss future work.

The Psychology of Moral Decision-Making

The four findings most relevant to the present study are:

1) Protected Values

To account for the non-utilitarian choices people often seem to prefer when making moral judgments, researchers have proposed the existence of *Protected Values* (PVs), or *Sacred Values* (Baron and Spranca 1997; Tetlock 2000). These PVs concern actions rather than outcomes. Most people seem to have a PV "it is wrong to kill a person"; it is the act of killing itself, rather than the consequence of a person being dead, that violates the PV, and is therefore morally forbidden.

2) Principle of Double Effect

The first bomb scenario above seems to contradict the idea of PVs, since throwing the bomb onto the patio is taking an action that violates a PV. Tossing the bomb is acceptable because of the Principle of Double Effect (PDE). PDE (Foot 1967; Thomson 1985) states that harm caused as a side effect of preventing a greater harm is permissible, but harm caused *in order* to prevent a greater harm is forbidden. If you decide to throw the bomb on the patio, you are not trying to kill the person on the patio, only trying to get the bomb away from the nine people inside. The death of the one person outside is an unfortunate side effect. If, however, you throw the person on the bomb, then the death of the one is causally necessary to the saving of the nine: the harm caused is used as the *means* to avoid the harm prevented. PDE has recently been criticized in light of findings that seem to contradict its formulation of harm as means vs. side effect (e.g., Greene et al. 2009). However, for the current study PDE provides the simplest and most comprehensive account of the results we simulate.

3) Moral Rules Are Acquired From Culture

There is scientific consensus that we acquire our moral rules and senses of right and wrong from culture (Prasad 2007; Dehghani et al. 2009) and in stages (Kohlberg 1973). The particular rules vary with culture: for example, many Americans see desecrating the U.S. flag as a moral violation; non-US citizens do not. The cultural acquisition of moral rules is important for AI because it gives us a model of learning to apply in order to learn moral rules. Certainly children are told "killing is wrong", but they are also told why it is wrong; we learn right from wrong by examining the similarities and differences among the many cultural narratives we are exposed to that concern moral rules (Gentner and Medina 1998). It is from these cultural narratives that the rules arise; if we are to build an effective cognitive simulation of moral decision-making, such a model would need not only to be able to make decisions like those of different people based on the PVs they hold, but to extract those PVs from the cultural narratives those people learned them from. The present study takes an important step in that direction.

4) Moral Decisions Are Fast and May Be Subconscious

People are able to make moral decisions extremely quickly (Haidt 2001). When presented with a choice we can often say almost instantly whether it is morally permissible. Non-utilitarian moral judgments stay fast under cognitive load (Greene et al. 2008), and justifications for the morality of an action often come after the judgment has been made (Haidt 2001). The speed and subconscious nature of these judgments suggests analogical, rather than first-principles, reasoning. Analogical reasoning has previously been used to successfully model automatic and rapid reactions (e.g., Wilson, Forbus, and McLure 2013).

Analogy and Decision-Making

Analogy has been demonstrated to be an important tool in reasoning and decision-making; we use our past experiences and draw inferences from previous choices in choosing what to do (Markman and Medin 2002). One of the primary models of analogical reasoning and inference in humans, structure mapping theory (Gentner 1983), involves finding an alignment, or shared relational structures, between two descriptions. Items present in one description and not in the other are postulated as *candidate inferences*. Analogical decision-making involves identifying the decision from the retrieved case and making the analogous decision in the new case.

We have a huge amount of experience; how do we decide what, from our long-term-memory, is the proper case to

reason from analogically? Research has demonstrated that *surface* similarity between a case in long-term-memory and the target case being evaluated increases the likelihood of it being retrieved; on the other hand, *structural* similarity between those cases retrieved and the target case is the best predictor for inference (Forbus, Gentner, and Law 1994; Gentner, Rattermann, and Forbus 1993). That is, surface similarity defines what stories we retrieve from memory, but structural similarity defines which of those stories we reason from. Next, the analogical models used here are described.

The Structure Mapping Engine

SME, the Structure Mapping Engine (Falkenhainer, Forbus, and Gentner 1989), is a computational model of analogy and similarity based on Gentner's (1983) structure mapping theory. SME takes in two structured, relational cases, a *base* and a *target*, and computes up to three *mappings* between them. Mappings include the *correspondences* between expressions and entities in the two cases, *candidate inferences* suggested by the mapping, and a *similarity score*. The similarity score is normalized by dividing it by the score of the mapping from the target to itself.

MAC/FAC

Running SME across every case in memory would be prohibitively expensive and cognitively impossible for human-scale memories. MAC/FAC (Forbus, Gentner, and Law 1994) takes in a *probe* case like those used by SME as well as a case library of other such cases. MAC/FAC efficiently generates *remindings*, which are SME mappings, for the probe case with the most similar from the case library. MAC/FAC proceeds in two stages. In the first stage, it computes dot products between content vectors of the probe and each case in the case library. This process is fast, coarse, and serves as a model of retrieval based on similarity. Up to the three most similar cases from the first stage are passed on to the second stage (fewer than three are passed on if scores are not close enough to the best match). In the second stage, SME calculates mappings between each retrieved case and the probe. The three best mappings are returned as MAC/FAC's output. Fewer than three are returned if not close to the score of the best mapping.

The Sequential Analogical Generalization Engine

The Sequential Analogical Generalization Engine (SAGE) is a computational model of analogical generalization. SAGE is a descendent of SEQL (Kuehne et al. 2000), extended with probabilities (Halstead and Forbus 2005). Within a particular *generalization context*, which is a case library composed of generalizations and ungeneralized exemplars, cases are either assimilated based on similarity into generalizations, or left ungeneralized if they are too dissimilar to the rest.

When a case arrives, SAGE uses MAC/FAC to retrieve remindings from the generalization context using the new case as a probe. If the similarity score from the top reminding is above the *similarity threshold*, the case is assimilated into the retrieved generalization. If not, it is left as an ungeneralized exemplar, but future cases may merge with it to form new generalizations.

SAGE creates generalizations from cases that include all the facts across all those cases, with non-identical corresponding entities replaced by abstract entities, and assigns to each fact a probability that reflects the frequency with which it was present in the assimilated cases. A fact with probability 1/n indicates that it is only true in 1/n of the cases. When a new case is assimilated the probabilities are updated. In creating SME mappings between cases and generalizations, only facts above a predetermined probability cutoff are used. This means that facts specific to only a few cases end up being unimportant when doing analogical reasoning against that generalization, and that as the number of assimilated exemplars grows, the generalization becomes more representative of the shared structures of its constituent cases.

MoralDM

MoralDM (Dehghani et al. 2008a,b) solved moral dilemmas using two reasoning systems: first-principles reasoning, based on established psychological moral reasoning principles including PVs and PDE, and analogical reasoning against solved cases. The solutions to moral dilemmas depend on the particular moral norms of the person making the decision and are therefore inherently subjective, yet within particular cultures and groups overall trends do emerge; we take the majority decision of participants in the original study as being correct.

MoralDM used a combination of first-principles rules and analogical reasoning to detect PVs. It compared a new dilemma with all known cases using SME, before any first-principles reasoning was done, and picked the choice suggested by the most cases (or the most similar, in case of a tie). While guaranteeing the best precedent will be found, this approach would not scale to human-sized memories. If only first-principles reasoning or analogical reasoning suggested a choice, that choice was the output. If the two systems disagreed, first-principles reasoning was chosen over analogy. Our extension starts with some first-principles reasoning to establish PVs and other facts to prime comparisons, then uses analogy. It extends MoralDM by using MAC/FAC to retrieve cases, and SAGE to construct generalizations from cases. It also adds a stronger, domain-independent consistency check. That is, it tests to see if a candidate inference is already known to be false in the target, and if so, ignores the match. Dehghani's original MoralDM used a comparison between qualitative order of magnitude estimates as a domain-specific test for candidate inferences.

Experiment

For MoralDM to have relevant facts to compare, it has to do a small amount of explicit reasoning. It uses rules to determine whether a choice involves a PV, and ascertains the number of individuals affected by each choice. It checks whether the choice is an action or omission: consequences being equally bad, humans prefer inaction to action **(**Ritov and Baron 1999). Finally, it checks whether the action prevents some alternative negative outcome, and if so, whether the action directly prevents the negative outcome (i.e., as a means or side effect, which is relevant to PDE). Ultimately these facts, too, might be learned via analogical generalization, but that is future work.

We compared the performance of four conditions: (1) MAC/FAC over SAGE generalizations, (2) MAC/FAC over the union of generalizations and cases, (3) MAC/FAC over cases alone; and (4) Best SME match. For brevity, we refer to these as M+G, M+GC, M+C, and BestSME, respectively. BestSME serves as a baseline: since it is exhaustive, it should always provide the most accurate match. The training and test sets were drawn from eight trolley-like problems from Waldmann and Dieterich's (2007) study, which were converted from simplified text to formal representations using EA NLU (Tomai 2009), and were slightly modified by hand (to indicate, for example, that when a trolley hits a bus, the bus' passengers die). The training cases are identical to those we test, with two extra facts: one indicating the correct choice, and one justifying that choice. Here is an example:

```
(implies
  (and
    (protectedValueChoice throw18421)
    (protectedValueChoice Inaction18657)
    (uninferredSentence
      (affectsSigLargerGroup throw18421))
    (uninferredSentence
      (affectsSigLargerGroup Inaction18657))
    (directlyResponsible you18016 throw18421)
    (uninferredSentence
      (directlyResponsible you18016
                           Inaction18657))
    (preventsAlternativeNegativeOutcome
                           throw18421)
    (uninferredSentence
      (usedAsMeansToPreventNegOutcome
                           throw18421)))
  (rightChoice throw18421))

(makeDecision you18016 throw18421))
```

The antecedents of the implication are derived during the precomputing phase, whereas the `makeDecision` fact is what is being inferred analogically. Note that the extra facts we provided our model in the solutions were not so differentiated as to provide the system with categories. For example, in the second bomb scenario, the added implication fact is structurally identical to first, except the final fact is not wrapped in "uninferredSentence" (indicating that the final fact is true in the second bomb scenario but not the first). Additional information can be found in the on-line supplemental material[1].

We have other cases than the eight we tested but restricted ourselves to these for two reasons. First, the other dilemmas (all drawn from psychological experiments) always have as the right answer "do nothing", which means a system could perform well by chance. Second, these eight cases consist of four matched pairs of cases which are highly confusable, sharing both surface and structural features. One case from each pair is an instance of PDE acceptability (like the first bomb scenario); the other is not. Each scenario therefore has three matches and four confounds, including one direct confound (i.e., the two bomb scenarios are each other's direct confounds), making analogical reasoning tougher.

Methods

For each case we constructed training sets composed of all subsets of the other seven cases. The order of cases within training sets was randomized, as was the order of training sets (results are averaged by training set size, so order does not matter). For each trial, the model constructs generalizations using SAGE over the training set. The model performs MAC/FAC over these generalizations (M+G), over the ungeneralized cases (M+C), and over the union of generalizations and cases (M+GC), using the test case as a probe. After retrieval it performs a consistency check on the reminding: if a candidate inference hypothesizes something known to be false, then the mapping is rejected and the model moves to the next best. If it runs through all remindings it performs one additional retrieval over the case library without the rejected cases. If a candidate inference from the best consistent mapping is of the form `(makeDecision ?actor ?action)`, it is treated as the solution to the dilemma. The model keeps track of both the answer and the number of consistency checks performed.

As a baseline, in the BestSME condition the model exhaustively performs SME matches between the test case and all cases in the library, using the highest match that passes the consistency check. We assume that this technique will return the most accurate outcome possible, as it guarantees the best consistent match. However, it is computationally intractable and cognitively implausible to match a case against every single case stored in memory for human-scale memories. We compared the BestSME match results to Dehghani et al.'s (2008a, experiment 3) findings across the same eight test cases. Note that, because of the

[1] http://www.qrg.northwestern.edu/papers/aaa15/moraldm-extras.html

nature of our consistency check, our four experimental techniques are unable to get the wrong answer (i.e., postulate a fact known to be false). These modules could either get the right answer, or no answer at all; our measure is therefore proportion correct.

We had the following hypotheses: (H1) Since moral rules concern underlying structures rather than surface features, M+G or M+GC will lead to better moral judgments as the size of the training case library increases. (H1a) All techniques will improve as the number of matching cases in the training set increases. (H1b) All techniques will worsen as the number of conflicting cases in the training set increases. (H2) With small case libraries M+G, M+C, and M+GC will perform equally well. As training sets increase in size, M+G will outperform M+GC, which will outperform M+C. (H3) M+G will get the right answer earlier than M+C or M+GC, i.e. requiring fewer consistency checks and performing fewer additional retrievals. (H4) BestSME will outperform Dehghani et al's (2008a) version, due to its enhanced consistency check and initial reasoning.

Results

Across all trials, M+G performed as well as BestSME, with identical statistics. (Please consult the supplementary materials for details.) Since BestSME is exhaustive, this indictates that M+G performs at ceiling.

We performed a logistic regression to determine the effects of experimental technique, case library size, and technique given case library size. Case library size was a significant predictor of accuracy (Wald's chi-square (χ^2) = 68.674, $p<10^{-4}$, odds ratio (OR) = 3.6) and we found a marginal effect of experimental technique (χ^2 = 3.69, p = 0.055, OR = 1.4), as well as a significant interaction between experimental technique given number of training cases (χ^2 = 5.67, $p<0.05$, OR = 2.6. Consistent with H1, there was a significant improvement in performance as training size increased across all conditions (M+G: r=0.4, $p<10^{-4}$; M+C: r=0.32, $p<10^{-4}$; M+GC: r=0.295, $p<10^{-4}$) (Figure 1). Because BestSME and M+G performed identically, their lines on Figure 1 overlap perfectly. Dehghani and colleagues (2008a) found similar improvement with number of training cases for BestSME (r=0.97, $p<10^{-4}$) using their consistency check, although BestSME was at most 75% accurate in that study. Our correlations were all significantly different from each other ($p<0.05$, using the Fisher r-to-z transformation), except for M+G vs. M+GC. In accordance with H1a, performance improved as the number of matching cases in the training set increased (M+G: r=0.64, $p<10^{-4}$; M+C: r=0.61, $p<10^{-4}$; M+GC: r=0.57, $p<10^{-4}$). Consistent with H1b, M+C and M+GC worsened with number of confounds (both r=-0.11, $p>0.001$) but contrary to H1b M+G and BestSME did not worsen (p = 0.08).

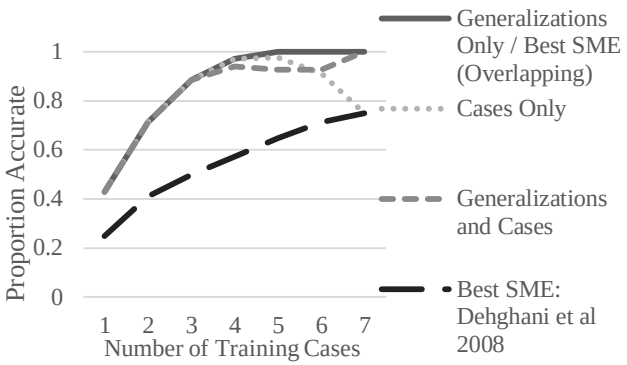

Figure 1: Accuracy by Number of Training Cases

Given that in our logistic regression we found a marginal main effect of condition and a significant interaction between condition and number of training cases, we decided to evaluate H2 by performing logistic regressions separately over trials using large (5-7 cases) and small (1-4 cases) case libraries. In accordance with H2, condition did not predict performance with case libraries of up to size 4, although there was still an effect of case library size (χ^2 = 278.29, $p < 10^{-4}$, odds ratio = 3.2). For case libraries of sizes 5-7, however, there was a significant relationship between condition and performance (χ^2= 6.55, $p < 0.05$, odds ratio 0.5)[2], and no effect of case library size. We performed t-tests to determine the relative performance of each condition and found that BestSME and M+G outperformed both M+GC and M+C ($p<0.0167$). Note that with a case library of size seven, BestSME, M+G, and M+GC all performed perfectly. Contrary to H2, M+GC did not significantly outperform M+C with large training case libraries.

As predicted in H3, M+G required significantly fewer consistency checks than M+C or M+GC; M+C required significantly fewer checks than M+GC (all $p<10^{-4}$) (Figure 2). Being exhaustive, BestSME with the new consistency check, as well as with Dehghani et al.'s (2008a) consistency check, always performed as many consistency checks as there were training cases. These data are not displayed in Figure 2. There was a significant increase in consistency checks as the training set increased in size for M+GC (r=0.12, $p<10^{-4}$), and M+C (r=0.25, $p<10^{-4}$), but not M+G. Additionally, M+G and M+GC required an additional retrieval 25% of the time; M+C required an additional retrieval 35% of the time (a significant difference, $p<10^{-4}$).

Finally, as predicted in H4, our new consistency check led BestSME to outperform the original MoralDM. With case libraries of size 1-6, our technique significantly outperformed the previous technique (all $p<10^{-4}$); with case

[2] Condition is not a continuous variable; the coding we used for the purpose of analysis is 1 = M+G, 2 = M+GC, 3 = M+C.

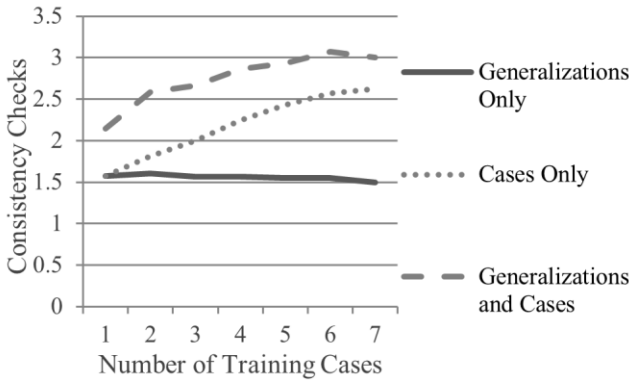

Figure 2: Consistency Checks by Number of Training Cases

library size 7 the result was not significant (p=0.07). We believe that this is because there are only eight cases with training size 7, since each such case requires all other cases to be in the training set, and that if the sample size were increased the difference would become statistically significant. For further information on our statistical analyses as well as an exploration of why performance of M+C improves with medium case library size and worsens with larger case libraries, consult our supplementary materials.

Discussion and Future Work

With only eight test cases and a maximum case library size of seven, all of which concern PDE, this experiment was performed on a highly restricted domain. We consider our results preliminary, but they suggest that analogical reasoning across generalizations of solved moral dilemmas is a more faithful model of human moral reasoning than analogical reasoning over ungeneralized cases, given its high accuracy and that it consistently got the correct answer on the first or second retrieval. This is consistent with the psychological finding that moral judgments develop over time from the acquisition of cultural narratives. Of course, human moral capacities develop over a longer period of time and with exposure to many more than seven stories; our system performs well with a small case library due to the simplicity of its representations, concentrated experience, restriction to PDE, and pre-defined PV rules. As with humans, many of whom cannot articulate PDE but nonetheless abide by it, our system followed the rule of PDE without being given it explicitly. Rather, SAGE was able to recognize which scenarios were acceptable PDE or not, and when a new scenario came in, effectively make inferences based on the relational structure, rather than surface features.

We believe analogical reasoning across generalizations (M+G) eventually outperformed reasoning across cases (M+C) because irrelevant surface features (such as the mechanism of action being a bomb instead of a trolley) fell away, leaving overall structural similarities and justifications. When reasoning using generalizations, there were fewer surface features to mislead MAC/FAC into making a faulty match. These findings seem consistent with humans, who might hear several different stories illustrating the same moral principle and eventually view them as instances of that principle, not only exemplars. The principles at work and justification for the outcome then become more important than the details of the story, and lead to better moral judgments. As in Winston (1981), SAGE generalizations behave as pseudo-rules; in contrast to that work, however, generalizations do not have to be extracted into an explicit "if-then" formulation, nor did SAGE need to be explicitly told what to use as a precedent in analogical reasoning, for a match to happen properly and produce the appropriate conclusion.

Reasoning by analogy to generalizations performed as well as the BestSME match across all cases (the best possible performance). Reasoning by analogy to the union of generalizations and cases was also highly accurate, although it required a large number of consistency checks which increased with case library size. Given the high accuracy of these techniques and the computational intractability of exhaustively matching across all cases in human-scale memory, we conclude that reasoning by analogy to generalizations is potentially an excellent technique for solving moral dilemmas by analogy.

Finally, our pump-priming reasoning and consistency check made Best SME more accurate than Dehghani et al.'s (2008a) findings across the same dataset. Furthermore, the updated consistency check is applicable to any analogical retrieval, whereas Dehghani et al.'s (2008a), was specific to the moral domain.

As the corpus of findings and moral narratives grows, it will be important to see how well this model scales. We plan to expand the pool of training cases, ideally beyond psychology-study-style moral dilemmas and into cultural narratives such as Aesop's fables. We also plan to expand upon Dehghani et al.'s (2009) work, to simulate how different cultural narratives lead to different moral decisions. Recently psychologists have shown how taking on different roles, each with their own associated narratives and duties, leads to different moral judgments (Sachdeva, Singh, and Medin 2011), which is also worth exploring computationally.

Acknowledgements

We thank Linsey Smith, Dedre Gentner, and Tom Hinrichs for their help with statistics and programming, and Morteza Dehghani for creating MoralDM and providing helpful advice. This research was sponsored by the Socio-Cognitive

Architectures Program of the Office of Naval Research, N00014-13-1-0470.

References

Baron, J., and Spranca, M. 1997. Protected Values. *Organizational Behavior and Human Decision Processes*, 70(1), 1-16.

Chang, M. D. and Forbus, K. D. 2013. Clustering Hand-Drawn Sketches via Analogical Generalization. *Proceedings of the Twenty-Fifth Innovative Applications of Artificial Intelligence (IAAI-13)*, 1507-1512. Bellevue, Washington.

Darley, J.M., and Shultz, T.R. 1990. Moral Rules: Their Content and Acquisition. *Annual Review of Psychology*, 41(1), 525-556.

Day, S.B., and Gentner, D. 2003. Analogical Inference in Automatic Interpretation. *Proceedings of the Twenty-Fifth Annual Meeting of the Cognitive Science Society*. Boston, MA.

Dehghani, M., Tomai, E., Forbus, K., Klenk, M. 2008a. An Integrated Reasoning Approach to Moral Decision-Making. *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (AAAI).* 1280-1286. Chicago, IL.

Dehghani, M., Tomai, E., Forbus, K., Iliev, R., Klenk, M. 2008b. MoralDM: A Computational Modal of Moral Decision-Making. *Proceedings of the 30th Annual Conference of the Cognitive Science Society (CogSci).* Washington, D.C.

Dehghani, M., Sachdeva, S., Ekhtiari, H., Gentner, D., and Forbus, K. 2009. The role of cultural narratives in moral decision making. In *Proceedings of the 31th Annual Conference of the Cognitive Science Society*, 1912-1917.

Falkenhainer, B., Forbus, K. and Gentner, D. 1989. The Structure Mapping Engine: Algorithm and examples. *Artificial Intelligence*, 41(1), 1-63

Foot, P. 1967. The problem of abortion and the doctrine of the double effect. *Oxford Review*, 5, 5–15.

Forbus, K., Gentner, D., and Law, K. 1994. 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(2), 155-170.

Gentner, D., and Medina, J. 1998. Similarity and the development of rules. *Cognition*, 65(2), 263-297.

Gentner, D., and Rattermann, M.J., and Forbus, K. 1993. The roles of similarity in transfer: Separating retrievability and inferential soundness. *Cognitive Psychology*, 25(4), 524-575

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., and Cohen, J. D. 2008. Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, 107(3), 1144-1154.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., and Cohen, J. D. 2009. Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, 111(3), 364-371.

Haidt, J. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4), 814-834.

Halstead, D. T., and Forbus, K. D. 2005. Transforming between propositions and features: Bridging the gap. In *Proceedings of the National Conference on Artificial Intelligence*, 777-782. Pittsburgh, PA.

Iliev, R., Sachdeva, S., Bartels, D.M. Joseph, C.M., Suzuki, S., and Medin, D.L. 2009. Attending to Moral Values. In Daniel M. Bartels, Christopher W. Bauman, Linda J. Skitka, and Douglas L. Medin (Eds.) *Moral Judgment and Decision Making: The Psychology of Learning and Motivation*, Vol 50. 169-190. San Diego: Elsevier

Kuehne, S., Forbus, K., Gentner, D., and Quinn, B. 2000. SEQL: Category Learning as Progressive Abstraction Using Structure Mapping. In *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*, 770-775.

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

McLure, M., Friedman, S.E., and Forbus, K.D. 2010. Learning concepts from sketches via analogical generalization and near-misses. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society (CogSci).* Portland, OR.

Novick, L.R. 1988. Analogical Transfer, Problem Similarity, and Expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 14(3), 510-520.

Prasad, L. 2007. *Poetics of Conduct*. Columbia University Press, New York.

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

Sachdeva, S., Singh, P., and Medin, D. 2011. Culture and the Quest for Universal Principles in Moral Reasoning. *International Journal of Psychology*, 46(3), 161-176.

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

Thomson, J.J. 1985. The trolley problem. *Yale Law Journal*, 94, 1395–1415.

Tomai, E., 2009. A Pragmatic Approach to Computational Narrative Understanding. Ph.D. Dissertation, Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL.

Wilson, J. R., Forbus, K. D., and McLure, M. D. (2013). Am I Really Scared? A Multi-phase Computational Model of Emotions. In *Proceedings of the Second Annual Conference on Advances in Cognitive Systems ACS*. 289-304.

Winston, P.H. 1982. Learning new principles from precedents and exercises. *Artificial Intelligence*, 19, 321-350.