Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Interactive Learning and Analogical Chaining for Moral and Commonsense Reasoning

Joseph A. Blass
Qualitative Reasoning Group, Northwestern University, Evanston, IL
joeblass@u.northwestern.edu

Abstract

Autonomous systems must consider the moral ramifications of their actions. Moral norms vary among people and depend on common sense, posing a challenge for encoding them explicitly in a system. I propose to develop a model of repeated analogical chaining and analogical reasoning to enable autonomous agents to interactively learn to apply common sense and model an individual's moral norms.

Challenge and Research Goals

Should a self-driving car put its passengers at risk and swerve to avoid a jaywalker, or protect its passengers and hit them? To participate in our society, computers need to share our ethics. As they become more autonomous, they must consider the moral ramifications of their actions. I intend to build an AI moral-reasoning system that strives for good, but can select amongst only bad options, by acquiring and applying human morals. This system will require basic common-sense reasoning capacities (e.g., to understand the difference between throwing rocks vs. foam balls). The system will use Repeated Analogical Chaining (RAC) to acquire and apply common sense knowledge in order to understand moral situations; will learn moral norms through natural-language interaction with humans and analogical generalization; and will apply these norms by analogy.

The range of moral norms and concerns (Graham et al., 2009) make hand-encoding individuals' morals or providing case-by-case instructions impossible. Users also may not have the technical skills nor understand their own morals enough to encode them themselves. Since human morals do not depend only on first-principles reasoning (FPR) (Haidt, 2001), and since moral rules contradict and trade off with each other, I intend to minimize FPR in the system. A FPR moral system would either need rules for all possible trade-offs or would freeze when moral obligations conflict. Analogical reasoning can avoid these problems if it can retrieve a good case to reason from.

MoralDM, SME, and Companions

MoralDM (Dehghani et al. 2009), the basis of my work, is a computer model of moral reasoning that takes in a moral dilemma in natural language, uses a natural language understanding (NLU) system to generate Cyc-derived predicate-logic representations of the dilemma, and uses analogy over resolved cases and FPR over explicit moral rules to make human-consistent moral decisions.

The Structure Mapping Engine (SME), based on Gentner's (1983) Structure Mapping Theory, aligns two relational cases and draws inferences from the alignment. SME can apply norms by analogy (Dehghani et al. 2009). Analogy is good for moral reasoning because morality is defined by the relationships between actors and events, not their features (e.g., the instrument of harm).

MAC/FAC is a two-step model of analogical retrieval. MAC computes in parallel dot-products between the content vectors of the probe and each case in memory (a fast, coarse similarity measure). FAC then performs SME mappings on the most similar cases. MAC sees cases with mostly the same entities as the probe as good potential matches, even if the structures differ.

The Sequential Analogical Generalization Engine (SAGE) builds case generalizations that emphasize shared, and deprecate case-specific, facts. SAGE uses a case library of generalizations and examples. Generalizations contain facts from constituent cases: non-identical corresponding entities are replaced by abstract ones; probabilities indicate the proportion of cases each fact is in. Given a probe, SAGE uses MAC/FAC to find the most similar case in its library. If the match is strong, the case is assimilated; if not, it is added as an example.

The Companion Cognitive Architecture emphasizes the ubiquity of qualitative representations and analogical reasoning in human cognition. Companions work interactively with humans (Forbus & Hinrichs, 2006).

Proposed Research and Progress

RAC will function by retrieving a relevant common-sense unit (CSU) of knowledge from a large case base to achieve a clearer picture of a case. For example, if considering a case where a brick was dropped on someone, RAC would let the system conclude first that the person was hurt, then that they became unhappy. I will also extend MoralDM in the Companion Architecture to learn to extract an individual's moral norms using SAGE. MoralDM will get natural language moral stories from a user, extract qualitative representations from them, and generalize over those. Given a new case, MoralDM will use RAC to understand the case, then apply learned morals by analogy. I will extend MoralDM's analogical reasoning, integrate emotional appraisal, and improve NLU for a moral lexicon.

Achievements

Previously MoralDM exhaustively matched over resolved cases, which is computationally expensive and cognitively implausible. SME over ungeneralized cases also sees feature-similar but morally-different cases as a good match. Using MAC/FAC over generalizations rather than examples mitigates this surface-similarity bias, since generalizations emphasize defining shared structures. We have found that reasoning by analogy over generalizations led to more human-like judgments than using ungeneralized cases (Blass & Forbus, 2015).

Work in Progress: to be complete by 2/16

Work is progressing on RAC for commonsense reasoning. We are selecting a subset of training questions from the COPA (Roemmele *et al.*, 2011) corpus of commonsense questions. We will encode CSUs relevant to solving these questions, then show that RAC can repeatedly find relevant CSUs and apply them in order to solve these questions. This will require further extensions to QRG's NLU system, EA NLU. Our goal is a system that is able to acquire and apply commonsense knowledge as needed, not a system that can robustly answer any commonsense question.

EA NLU generates qualitative representations from English text, but its moral vocabulary is limited. The Moral Foundations Dictionary (Graham et al., 2009) is a moral lexicon; to enable EA NLU to understand moral tales, I am providing lexical and ontological support for this lexicon.

Future Work & Thesis

Developing the next generation MoralDM system will involve incorporating other QRG systems. Reasoning will benefit from McLure & Forbus' (2015) work on near-misses to illustrate category boundaries and conditions for membership or exclusion. Near-misses will help segment and define the boundaries of CSUs and moral norms. I will also integrate emotional appraisal (Wilson et al. 2013) into MoralDM, to help recognize violations and enforce decisions. We will also continue to add CSUs as needed.

We want to expand the range and source of stories for MoralDM to learn from. I have begun generating stories to teach MoralDM about morally charged situations, like revenge. I will investigate crowd-sourcing moral stories for a user to endorse or reject, to lighten the user's burden.

My thesis goal is to have a Companion running MoralDM with the above extensions interact with a human to build a model of their moral system. This was impossible when MoralDM required all morals to be explicitly encoded, and modeled societal, not individual, morality. The new system will have the human tell it a moral story, use RAC with CSUs to fully understand it, crowd-source thematically similar stories, and ask the human which illustrate the same moral principle (the others are near-misses). For each story, the system would predict the morality of actions and compare its predictions to the human's labels. When the core facts of a generalization stop changing and the system's labels consistently match the human's, the system has mastered that moral domain.

This project brings challenges. Can SAGE capture moral subtleties? How must I extend EA NLU for moral stories? How can appropriate CSUs be retrieved to infer implicit information? Nonetheless, I believe I can build a system that interactively learns to model an individual's morals.

References

Blass, J. & Forbus, K. (2015). Moral Decision-Making by Analogy: Generalizations vs. Exemplars. *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, Austin, TX.

Dehghani, M., Sachdeva, S., Ekhtiari, H., Gentner, D., & Forbus, K. (2009). The role of cultural narratives in moral decision making. *Procs of the 31st Annual Conf. of the Cog. Sci. Society.*

Forbus, K., & Hinrichs, T. (2006). Companion Cognitive Systems: A Step Towards Human-Level AI. *AI Magazine* 27(2).

Gentner, D. (1983). Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Science* 7(2).

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5).

Haidt, J. (2001). The Emotional Dog & its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psych. Review*, 108(4).

McLure, M., Friedman S. and Forbus, K. (2015). Extending Analogical Generalization with Near-Misses. *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, Austin, TX

Roemmele, M., Bejan, C., and Gordon, A. (2011) Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, Stanford University

Wilson, J., Forbus, K., & McLure, M. (2013). Am I Really Scared? A Multi-phase Computational Model of Emotions. *Procs of the 2nd Annual Conf. on Advances in Cognitive Systems.*