Interactively Learning Moral Norms via Analogy

Joseph Blass

Ph.D. Candidate, 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, posing a challenge for encoding them explicitly in a system. This paper proposes to enable autonomous agents to use analogical reasoning techniques to interactively learn an individual's morals.

1 Introduction

1.1 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 him? To participate in our society, computers need to share our ethics. As these systems 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 learn moral norms through natural-language interaction with humans and analogical generalization, and apply these norms by analogy.

The diversity of moral norms and concerns make hand-encoding an individual's moral sense or providing case-by-case instructions impossible. Natural interaction will be key, since users may have neither the technical skills nor understand their own morals enough to encode them themselves. Also, since human morals likely do not depend 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 pure FPR moral reasoning system would either need rules for all possible trade-offs, to be able to ignore certain morals (a bad idea), or would freeze when moral obligations conflict. Analogical reasoning can avoid these problems if provided a good analogue.

1.2 MoralDM, Structure-Mapping, and the Companions Architecture

MoralDM (Dehghani et al. 2009) 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 Research-Cyc-derived predicate-logic representations of the dilemma, and uses analogy over resolved cases and FPR over explicit moral rules to make moral decisions consistent with humans'. MoralDM is the starting point for my work.

The Structure Mapping Engine (SME), based on Gentner's (1983) Structure Mapping Theory of analogy, constructs an alignment between two relational cases and

draws inferences from it. SME can apply norms by analogy from stories (Dehghani et al. 2009). Analogy is a good fit for moral decision-making because both are guided by structure, not features. Consider the following examples. 1) A bomb will kill nine people in a room, but you can toss it outside, where it will kill one person. 2) A bomb will kill nine people, but you can toss someone onto it to absorb the blast and save the nine. Most say tossing the bomb, but not the person, is morally acceptable. These scenarios only differ structurally, in what fills which role; the entities and action types themselves are shared. The classic trolley problem (a trolley will hit five people unless it is diverted to a side track where it will hit one person), in contrast, has different features, but the same structure, as the first bomb case. Humans see these two cases as morally alike.

The Sequential Analogical Generalization Engine (SAGE) builds case generalizations that emphasize shared, and deprecate case-specific, structures. SAGE uses a case library of generalizations and exemplars. Generalizations contain facts from constituent cases: non-identical corresponding entities are replaced by abstract ones; probabilities indicate the proportion of assimilated cases each fact is present in. Given a probe, SAGE uses SME to find the most similar case in its case library. If the match is strong enough, the case is assimilated; if not, it is added as an exemplar. SAGE can use near-misses to determine defining characteristics of category members (McLure et al., 2015).

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

2 Proposed Research and Progress

I propose to extend MoralDM in the Companion Architecture to learn to model a human user's morals. The system will learn to recognize and extract moral norms through the generalization process. It will get moral stories in natural language from the user, generate qualitative representations of those stories, generalize over those representations, and use SME to apply morals from the generalizations. I will extend MoralDM's analogical reasoning, integrate emotional appraisal, and improve NLU for a moral lexicon.

Previously MoralDM's analogical reasoning module exhaustively matched over resolved cases, which is computationally expensive and cognitively implausible. SME over ungeneralized cases also sees feature-similar but morally-different cases (i.e., the bomb scenarios) as a good match, due to the amount they have in common.

MAC/FAC is a two-step model of analogical retrieval. MAC efficiently computes dot-products between the content vectors of the probe and each case in memory (a coarse similarity measure). FAC then performs SME mappings on the most similar cases. MAC sees cases concerning mostly the same entities as the probe as good potential matches, even if the structures differ. Using MAC/FAC over generalizations rather than exemplars solves this problem, since generalizations emphasize defining structure. Abstract generalizations applied by analogy can therefore function as moral rules.

We have found that reasoning by analogy over generalizations led to more human-like judgments than using ungeneralized cases (Blass & Forbus, 2015). Reasoning can be further improved using McLure & Forbus' (2015) work on near-misses to illustrate category boundaries and the conditions for membership or exclusion. MoralDM also

still reasons using FPR about facts relevant to moral judgment, such as directness of harm. These are not explicitly stated, though we recognize them easily; MoralDM uses them in a consistency check to ensure the quality of retrieved analogues. Near-misses would let MoralDM use analogy, not FPR, to find the facts for the consistency check.

We want to expand the range and provenance of stories for MoralDM to learn from. One option is to crowd-source moral stories to present to a user for endorsement or rejection, rather than force the user to provide them all. QRG's NLU system, EA NLU, generates qualitative representations from English input, but its moral vocabulary is currently limited. The Moral Foundations Dictionary (Graham et al., 2009) is a moral lexicon; to enable EA NLU to understand moral stories, I will ensure lexical and ontological support for this vocabulary. Another NLU challenge is how to infer information implicit in the text. Work has been done at QRG on inferring narrative information, including about moral responsibility (Tomai & Forbus, 2008). I will extend EA NLU's abductive reasoning as needed to support moral narrative understanding. Finally, I will integrate emotional appraisal (Wilson et al. 2013) into MoralDM. Emotional appraisal can help recognize moral violations and enforce moral decisions.

My goal is to have a Companion running MoralDM with the above extensions interact with a human and build a model of their moral system. MoralDM could not previously do this, since it required all moral norms to be explicitly encoded, and modeled a society's aggregate judgments, not individuals. The new system will have the human tell it a moral story, 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 moral value of actions and compare its predictions to the human's moral labels. When the core facts of the generalization stop changing and the system's labels consistently match the human's, the system has mastered that moral domain.

This project brings challenges. How much FPR will remain necessary? How must EA NLU be extended to understand moral narratives? What narrative inferences should be made about implicit information? Nonetheless, I believe I can build a system that interactively learns to model an individual's morality.

3 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. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Forbus, K., & Hinrichs, T. (2006). Companion Cognitive Systems: A Step Towards Human-Level AI. *AI Magazine* 27(2), 83-95.

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), 1029.

Haidt, J. (2001). The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review*, 108(4), 814-834.

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

Tomai, E. & Forbus, K. (2008). Using Qualitative Reasoning for the Attribution of Moral Responsibility. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*. Washington, D.C.

Wilson, J. R., Forbus, K. D., & 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*.