Learning and obeying conflicting norms in stochastic domains

Daniel Kasenberg
Tufts University
200 Boston Ave, Suite 2600
Medford, MA 02155

Introduction

Artificial agents will need to be able to reason about and obey human moral and social norms. Additionally, agents must be able to *learn* norms, both from instruction (e.g., by natural language interaction with humans) and by observing the behavior of humans or other agents. This is necessary because (1) humans have many moral and social norms, perhaps too many to be able to pre-program; and (2) moral and social norms may vary between cultures, and across time.

Both of these challenges – obeying moral and social norms, and learning them – are complicated by the fact that such norms may conflict. When an agent's norms cannot all be satisfied, that agent should be able to "minimize the badness" of their behaviors in some principled way. Norm conflicts also make *learning* norms (e.g., from behavior) more difficult, since (1) the agent will need to learn preferences between norms as well as the norms themselves (in case of future conflict), and (2) the observed behaviors may be a compromise between inconsistent normative principles.

Our work addresses the problem of learning (from behavior) and obeying potentially-conflicting norms in stochastic domains. In particular, we represent moral and social norms as statements in Linear Temporal Logic (LTL), and we employ the framework of Markov Decision Processes (MDPs).

Related work

We compare our work to two particular approaches to learning and obeying moral and social norms: those involving *logic*, and those involving *reward*.

Many of the "logical" approaches to conflicting norms employ deontic logics. Some (Vasconcelos, Kollingbaum, and Norman 2009) modify (or "curtail") the norms themselves to ensure that conflicting norms never apply in the same contexts. Other approaches use nonmonotonic logics that avoid some of the problems associated with conflicting obligations (Beirlaen, Straßer, and Meheus 2013). These methods generally assume deterministic environments, and are not easily adapted to stochastic domains.

Reward-based approaches to AI ethics often use some variant of *inverse reinforcement learning* (IRL) (Ng and Russell 2000). These approaches seek to learn a reward function (presumably representing moral and social norms) from observed agent behavior. We have argued elsewhere (Arnold, Kasenberg, and Scheutz 2017) that such approaches may be incapable of representing temporally complex norms, and lack interpretability.

Linear temporal logic

Linear Temporal Logic (LTL) is a propositional logic augmented with temporal operators $\mathbf{X}, \mathbf{G}, \mathbf{F}, \mathbf{U}$. $\mathbf{X}\phi$ means "in the next time step, ϕ"; $\mathbf{G}\phi$ means "in all present and future time steps, ϕ"; $\mathbf{F}\phi$ means "in some present or future time step, ϕ"; and $\phi_1 \mathbf{U} \phi_2$ means "ϕ_1 will hold until ϕ_2 holds".

Results

Norm conflict resolution

Let $\Phi = (\phi_1, \cdots, \phi_N)$ be a set of LTL statements, and $\mathbf{w} = (w_1, \cdots, w_N)$ be a vector of nonnegative real weights. We may then define a *norm system* $\mathcal{N} = (\Phi, \mathbf{w})$ Here each weight w_i represents the relative importance the agent ascribes to the norm ϕ_i.

Ideally, the goal of a norm-obeying agent would be to find some (in general non-stationary) policy that obeys all LTL norms with probability 1. However, in practice this may not be possible. We call a *norm conflict* any situation in which the probability of an agent simultaneously obeying all of its norms is zero. We refer to the problem of determining a "least bad" course of action in the event of norm conflicts as *norm conflict resolution* (NCR).

In (Kasenberg and Scheutz 2018) we provide an NCR algorithm. To do so, we define a notion of "violation cost" as follows:

Considering an infinite behavior trajectory $\tau = s_0, a_0, s_1, a_1, \cdots$, we may "omit" some set J of time steps from τ to yield a modified trajectory $\tau \backslash J$. For example, if $N = \{1, 2\}$ then $\tau \backslash J = s_0, a_0, s_3, a_3, s_4, a_4, \cdots$. We can then define the *violation cost* $\mathrm{Viol}_\phi(\tau)$ of τ with respect to some norm ϕ as the (discounted) size of the smallest set of time steps J that must be omitted in order for $\tau \backslash J$ to satisfy ϕ. The violation cost of a trajectory τ with respect to a norm system $\mathcal{N}$ is weighted sum of the violation costs of τ with respect to each of the norms in $\mathcal{N}$.

The goal of the agent, then, is to find a (general) policy that minimizes *expected* violation cost with respect to the

norm system. To do this, we construct from each norm a corresponding deterministic Rabin automaton (DRA), a type of finite state machine over infinite words. We then construct a *product MDP* - the Cartesian product of the original MDP and the DRAs for each of the agent's norms. While the optimal expected violation cost is not Markovian in the original MDP, it *is* Markovian in the product MDP, and can thus be computed by value iteration (with some graph-theoretic caveats; see (Kasenberg and Scheutz 2018) for details). This can be used to find the optimal policy (which is stationary in the product MDP).

In each of the simulated domains in which we tested our NCR algorithm, the results matched our intuition about the right behavior given the norms and their relative importance.

Norm inference

In (Kasenberg and Scheutz 2017), we developed a *norm inference* algorithm which, given some set of (finite) trajectories representing agent behavior, seeks to determine a temporal logic statement that "best explains" those norms.

We formulated norm inference as a multi-objective optimization problem $\min_{\phi \in LTL}(\text{Obj}^S(\phi), \text{Obj}^X(\phi))$, where Obj^S captures a notion of formula complexity. Obj^X represents the idea that norms that *specifically* explain the observed behavior (as opposed to random behavior) should be preferred:

$$\text{Obj}^X(\phi) = \text{Viol}_\phi(\pi^o) - \text{Viol}_\phi(\pi^{rand}) \quad (1)$$

where the $\text{Viol}_\phi(\pi)$ is the expected violation cost under the (product-space) policy π, π^o is the "observed" product-space policy determined by the observed trajectories, and π^{rand} is the uniformly random policy.

This problem may be solved using any multi-objective optimization algorithm capable of optimizing over a grammar. The result will be a set of Pareto-efficient solutions, from which a norm may be selected.

We have tested this approach in several simple domains, and found it was able to retrieve norms that explained the observed trajectories. We have also (Kasenberg, Arnold, and Scheutz under review) undertaken a comparison between our norm inference approach and reward-based approaches to learning from behavior.

Inverse norm conflict resolution

For agents attempting to learn potentially-conflicting norms by observing behavior, one crucial problem is to learn a set of relative preferences among these norms. To this end, in (Kasenberg and Scheutz under review) we define an "inverse norm conflict resolution" (INCR) algorithm that, given a set of behavior trajectories and a set Φ of norms, determines a vector $\mathbf{w}$ of weights which "best explain" the behaviors.

INCR proceeds by minimizing the relative entropy (or KL divergence) between the "observed product-space policy" π^o and the optimal product-space policy (the policy minimizing violation costs, determined by our NCR algorithm) with respect to the norms. The algorithm uses fixed-point iteration and gradient descent to minimize this objective; the result is a set of weights $\mathbf{w}^*$ so that the optimal norm-obeying policy for the norm system $(\Phi, \mathbf{w}^*)$ approximates the observed behavior.

Future work

Whereas our norm inference algorithm returns a set of Pareto-efficient norms such that only one should be chosen, future work will incorporate this with INCR to allow learning a *set of norms* (and a corresponding weight vector) that *jointly* explain the observed behaviors.

One advantage of explicitly representing norms in logic is their *interpretability*. This may enable the agent to learn norms through natural-language instruction, and to generate explanations or justifications when questioned about behavior in NCR scenarios. We aim to incorporate these capabilities in future work.

LTL is not sufficient to perform complex reasoning about duties, since it lacks explicit deontic operators. Future work may explore using temporal deontic logics that enable such reasoning. To do this, we would need to adapt our existing algorithms to such logics.

Conclusion

Through our work in norm conflict resolution, norm inference, and inverse norm conflict resolution, we have taken first steps towards our goal of an artificial agent able to learn and obey conflicting moral and social norms in stochastic domains.

References

Arnold, T.; Kasenberg, D.; and Scheutz, M. 2017. Value alignment or misalignment–what will keep systems accountable? In *3rd International Workshop on AI, Ethics, and Society*.

Beirlaen, M.; Straßer, C.; and Meheus, J. 2013. An inconsistency-adaptive deontic logic for normative conflicts. *Journal of Philosophical Logic* 1–31.

Kasenberg, D., and Scheutz, M. 2017. Interpretable apprenticeship learning with temporal logic specifications. In *Proceedings of the 56th IEEE Conference on Decision and Control (CDC)*.

Kasenberg, D., and Scheutz, M. 2018. Norm conflict resolution in stochastic domains. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

Kasenberg, D., and Scheutz, M. under review. Inverse norm conflict resolution. In *Proceedings of the First AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*.

Kasenberg, D.; Arnold, T.; and Scheutz, M. under review. Norms, rewards, and the intentional stance: Comparing machine learning approaches to ethical training. In *Proceedings of the First AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*.

Ng, A. Y., and Russell, S. J. 2000. Algorithms for inverse reinforcement learning. In *Icml*, 663–670.

Vasconcelos, W. W.; Kollingbaum, M. J.; and Norman, T. J. 2009. Normative conflict resolution in multi-agent systems. *Autonomous agents and multi-agent systems* 19(2):124–152.