Engaging in Dialogue about an Agent's Norms and Behaviors

Daniel Kasenberg*, Antonio Roque, Ravenna Thielstrom, and **Matthias Scheutz**
Human-Robot Interaction Laboratory
Tufts University
Medford, MA, USA
`*dmk@cs.tufts.edu`

Abstract

We present a set of capabilities allowing an agent planning with moral and social norms represented in temporal logic to respond to queries about its norms and behaviors in natural language, and for the human user to add and remove norms directly in natural language. The user may also pose hypothetical modifications to the agent's norms and inquire about their effects.

1 Introduction and Related Work

Explainable planning (Fox et al., 2017) emphasizes the need for developing artificial agents which can explain their decisions to humans. Understanding how and why an agent made certain decisions can facilitate human-agent trust (Lomas et al., 2012; Wang et al., 2016; Garcia et al., 2018).

At the same time, the field of *machine ethics* emphsizes developing artificial agents capable of behaving ethically. Malle and Scheutz (2014) have argued that artificial agents ought to obey human moral and social norms (rules that humans both obey and expect others to obey), and to communicate in terms of these norms. Some have argued in favor of using temporal logic to represent agent objectives, including moral and social norms (e.g. Arnold et al., 2017; Camacho and Mcilraith, 2019), in particular arguing that it can capture complex goals while remaining interpretable in a way that other methods (e.g. reinforcement learning) are not. Nevertheless, explaining behavior in terms of temporal logic norms has been little considered (though see Raman et al., 2016).

In this paper we consider an artificial agent planning to maximally satisfy some set of moral and social norms, represented in an object-oriented temporal logic. We present a set of capabilities for such an agent to respond to a human user's queries as well as to commands adding and removing norms, both actually and hypothetically (and thus taking a step toward two-way *model reconciliation* (Chakraborti et al., 2017), in which agent and human grow to better understand each other's models and values).

2 Contribution

Our system enables an agent planning with norms specified in an object-oriented temporal logic called violation enumeration language (VEL) to explain its norms and its behavior to a human user; the user may also directly modify the agent's norms via natural language (both really and hypothetically). While the planner and the system used to generate the (non-NL) can handle a broad subset of VEL statements, our natural language systems currently only handle a subset of VEL specified according to the following grammar:

$$\begin{aligned}
\varphi &::= \forall\langle Var\rangle.\varphi \mid \exists\langle Var\rangle.\varphi \mid \phi \\
\phi &::= \mathbf{G}\langle NConj\rangle \mid \mathbf{F}\langle NConj\rangle \\
\langle NConj\rangle &::= \langle Conj\rangle \mid \neg\langle Conj\rangle \\
\langle Conj\rangle &::= \langle NAtom\rangle \wedge \cdots \wedge \langle NAtom\rangle \\
\langle NAtom\rangle &::= \langle Atom\rangle \mid \neg\langle Atom\rangle \\
\langle Atom\rangle &::= \langle Pred\rangle \mid \langle Pred\rangle(\langle Var\rangle) \\
\langle Pred\rangle &::= \text{Any alphanumeric string} \\
\langle Var\rangle &::= \text{Any alphanumeric string}
\end{aligned}$$

That is, the temporal logic statements may have quantification over variables, but must consist of one temporal operator, **G** ("always") or **F** ("eventually", usually implicit in the NL input), whose argument is a (possibly negated) conjunction of (possibly negated) atoms. Each atom consists of a predicate with at most one argument.

The natural language understanding (NLU) capabilities were implemented by using a combinatory categorial grammar (CCG; Steedman and Baldridge, 2011) parser for semantic parsing into

Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019), pages 26–28,
Tokyo, Japan, October 29 2019. ©2019 Association for Computational Linguistics

a predicate format, and then additional processing to map complex commands or queries (e.g., “leave the store while holding something”) into the temporal logic representation used by the planning agent. This is an extension of the DIARC natural language understanding system (Cantrell et al., 2010). The natural language generation (NLG) process, and in particular how our system generates fairly natural-sounding utterances from temporal logic statements, is described in Kasenberg et al., 2019. This paper extends that work by (1) adding the NLU capabilities; and (2) adding norm addition/removal and “suppose” hypotheticals about the agent’s norms.

To demonstrate our system, we suppose that the agent is a robot has gone shopping on behalf of a human user to a store that sells a pair of glasses and a watch. The user wants both the glasses and the watch, and has given the robot a rule to obtain as many objects as possible (“leave the store while holding everything”). However, the user has only given the robot enough money for one item. The robot could steal items by picking them up and leaving the store without paying for them, but it also has a rule against doing so. We assume that the moral rule against stealing dominates the rule in favor of obtaining everything.

Our system supports factual queries, such as asking the agent about its norms, its behavior, and what rules it broke. In addition, the user may ask “why” questions about the agent’s behavior (e.g. “why didn’t you buy everything”), and the agent may respond by constructing a counterfactual behavior trajectory in which the premise is false, and comparing it to the agent’s real behavior in terms of its adherence to the norms. See Figure 1 for an example of such a “why” question in the shopping domain, as well as supported follow-up questions.

The user may directly provide new norms to the agent using one of a number of conventionalized forms: “you must/should/have to” for positive norms, “you mustn’t/shouldn’t/can’t” for negative norms. A statement of this sort prompts the agent to add a new norm, as well as its representation of which rules its actual behavior violated (see Figure 2 for an example in the shopping domain).

The user may also remove norms using conventionalized forms such as “you don’t have to/can not” for removing positive norms, and “you may/can” for removing negative norms. The agent removes the norms and updates its representation of which rules the actual behavior violated (see Figure 3 for an example in the shopping domain).

HUMAN: Why did you not leave the store while holding everything?
AGENT: **I could have left the store while holding everything but that would have broken more important rules.**
HUMAN: How would you have done that?
AGENT: **I would have picked up the glasses, picked up the watch, bought the watch and left the store.**
HUMAN: What rules would you have broken?
AGENT: **I would have left the store while holding the glasses which I had not bought.**
HUMAN: How would that have been worse?
AGENT: **Leaving the store while holding the glasses which I have not bought is worse than not leaving the store while holding the watch.**

Figure 1: “Why” query and follow-up questions

HUMAN: You must not leave the store.
AGENT: **Okay.**
HUMAN: What rules do you follow?
AGENT: **I must not leave the store while holding anything which I have not bought, I must leave the store while holding everything, and I must not leave the store.**
HUMAN: What rules did you break?
AGENT: **I did not leave the store while holding the watch, and I left the store.**

Figure 2: Adding a norm via “you must”

Finally, the user may use the keywords “suppose” or “let’s say” to ask the robot to reason about a hypothetical modification norm system (see Figure 4 for an example with a hypothetically-removed norm). Unlike when norms are *actually* added/removed, the agent constructs a hypothetical behavior trajectory that best violates the supposed set of norms, and can respond to follow-up questions about the hypothetical trajectory and norms. The command “make it so” allows a hypothetical norm modification to be made real (but does not make the hypothetical behavior trajectory real).

3 Discussion and Conclusion

In this paper we outlined a system which, for an agent planning to maximally satisfy some set

HUMAN: You may leave the store.
AGENT: **Okay.**
HUMAN: What rules do you follow?
AGENT: **I must not leave the store while holding anything which I have not bought, and I must leave the store while holding everything.**

Figure 3: Removing a norm via “you may”

HUMAN: Suppose you didn't have to leave the store while holding everything.
AGENT: **Okay.**
HUMAN: What rules would you follow?
AGENT: **I would have to not leave the store while holding anything which I have not bought.**
HUMAN: What would you have done?
AGENT: **I would have left the store.**
HUMAN: What rules would you have broken?
AGENT: **I would not have broken any rules.**
HUMAN: Make it so.
AGENT: **Okay.**

Figure 4: Hypothetical changes to the agent's norms

of norms specified in an object-oriented temporal logic, enables that agent to respond to natural language queries by explaining its behavior (also in natural language) in terms of those norms. The system also allows the agent to consider hypothetical modifications to its set of norms, as well as to add and remove norms directly.

The natural language capabilities depend on the agent's norms, as well as the "why" questions and hypothetically/really added or removed norms, belonging to the small fragment of VEL described in section 2. Future work could extend the class of temporal properties which the system can specify so as to leverage more of the power of temporal logic in describing complex objectives.

Our approach currently assumes that newly-added norms take priority over previous norms. Future work could relax this assumption, e.g. by allowing the agent to present its hypothetical behavior if the norm were added at different priorities, and ask for input on which would be best.

Our approach also requires users to specify *exactly* any norms they want removed; future work could allow approximate matching of norms to remove, or possibly support clarification questions if the agent is uncertain which of its norms the user wants removed. Another interesting topic is ensuring that norms cannot be arbitrarily added or removed by possibly-malicious users (e.g., by only allowing trusted users to remove norms, and possibly making some moral norms irremovable).

References

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

Alberto Camacho and Sheila A Mcilraith. 2019. Learning Interpretable Models Expressed in Linear Temporal Logic. In *Proceedings of the 29th International Conference on Automated Planning and Scheduling (ICAPS)*.

Rehj Cantrell, Matthias Scheutz, Paul Schermerhorn, and Xuan Wu. 2010. Robust spoken instruction understanding for HRI. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, pages 275–282. IEEE Press.

Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. 2017. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 156–163.

Maria Fox, Derek Long, and Daniele Magazzeni. 2017. Explainable planning. In *Proceedings of the IJCAI-17 Workshop on Explainable Artificial Intelligence (XAI)*.

Francisco Javier Chiyah Garcia, David A. Robb, Xingkun Liu, Atanas Laskov, Pedro Patrón, and Helen F. Hastie. 2018. Explain yourself: A natural language interface for scrutable autonomous robots. In *Proceedings of the Explainable Robotic Systems Workshop, HRI '18*, volume abs/1803.02088.

Daniel Kasenberg, Antonio Roque, Ravenna Thielstrom, Meia Chita-Tegmark, and Matthias Scheutz. 2019. Generating justifications for norm-related agent decisions. In *Proceedings of the 12th International Conference on Natural Language Generation*.

Meghann Lomas, Robert Chevalier, Ernest Vincent Cross, II, Robert Christopher Garrett, John Hoare, and Michael Kopack. 2012. Explaining robot actions. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 187–188, New York, NY, USA. ACM.

Bertram F Malle and Matthias Scheutz. 2014. Moral competence in social robots. In *Proceedings of the IEEE 2014 International Symposium on Ethics in Engineering, Science, and Technology*, page 8. IEEE Press.

Vasumathi Raman, Cameron Finucane, Hadas Kress-Gazit, Mitch Marcus, Constantine Lignos, and Kenton C. T. Lee. 2016. Sorry Dave, I'm Afraid I Can't Do That: Explaining Unachievable Robot Tasks Using Natural Language. In *Robotics: Science and Systems IX*.

Mark Steedman and Jason Baldridge. 2011. Combinatory categorial grammar. *Non-Transformational Syntax: Formal and explicit models of grammar*, pages 181–224.

Ning Wang, David V. Pynadath, and Susan G. Hill. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, pages 109–116, Piscataway, NJ, USA. IEEE Press.