Reasoning and Planning with Dynamic Social Norms

Extended Abstract

Taylor Olson
Northwestern University
Evanston, United States
taylorolson@u.northwestern.edu

Roberto Salas-Damian
Northwestern University
Evanston, United States
roberto.salas@u.northwestern.edu

Kenneth D. Forbus
Northwestern University
Evanston, United States
forbus@northwestern.edu

ABSTRACT

To safely interact with humans, AI systems must both have knowledge of our norms and consider norms in their planning processes. However, norm-guided planning has been less explored, only within communities of artificial agents and ignoring the dynamic nature of norms. This paper presents an approach to guiding planning with dynamically changing norms in a human-AI setting. This yields adaptive guard rails for the actions of AI systems.

KEYWORDS

Social Norms, AI Planning, Norm Learning and Reasoning, Privacy, Defeasible Reasoning

ACM Reference Format:
Taylor Olson, Roberto Salas-Damian, and Kenneth D. Forbus. 2025. Reasoning and Planning with Dynamic Social Norms: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

Imagine that Karli has told you "*Do not* share my medical records." Suppose Karli then gets married. So, she says, "*You may* tell my husband what prescriptions I'm taking." Now, suppose Karli then has children. So, she tells you, "*You must* share my health conditions with my children." How can an AI system respect Karli's dynamically changing wishes?

As AI systems become more integrated into our social world, they must be able to learn our norms and adapt their behavior accordingly. But while there have been developments formalizing norm learning and reasoning [7, 8, 10, 11], norm-guided planning has been less explored, only within communities of artificial agents and ignoring the dynamic nature of norms. Other approaches like LLMs may hold sophisticated dialogues at times, but they have proven to be manipulable and lack a solid theoretical foundation for normative reasoning.

This paper presents an approach to guiding plans with dynamically changing norms in a human-AI setting. We first provide background on the formal representations we draw upon. Next, we introduce our approach to reasoning about dynamically changing norms and utilizing them during planning. We conclude with an example and a discussion.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

2 BACKGROUND

2.1 NextKB Ontology

Our formalism utilizes the predicate calculus language of the NextKB[1] ontology. NextKB is derived from OpenCyc (an open-source subset of the Cyc ontology [4, 5]) and contains many concepts, relations, and facts. Knowledge in NextKb is contextualized using Cyc-style *microtheories* [2], enabling it to keep track of the beliefs of different agents.

2.2 Norm Frames

For representing norms, we draw upon the formalism of [7, 8]. This utilizes frame representations for both norms and actions (neo-Davidsonian action representations [1]) and thus better supports incremental learning than others, e.g. [12].

Definition 2.1 (Norm Frame). A *Norm Frame* is a logical encoding of a norm of the form:

```
(isa <norm> Norm)
(context <norm> <context>)
(behavior <norm> <behavior>)
(evaluation <norm> <deontic>)
```

Where `<norm>` is a constant representing the norm, `<context>` is a conjunction of literals that must be true for the norm to be *active*, `<behavior>` is a conjunction of positive literals representing the behavior that the norm *applies to*, and `<deontic>` is a modal of deontic logic [6]: `{Obligatory, Optional,Impermissible}`.

Definition 2.2 (Normative Belief). A normative belief is a particular agent's belief in a norm. For example, Karli believes that one should not share her medical records.

Definition 2.3 (Normative Testimony). Normative testimony is a natural language statement that introduces a norm [3]. For example, "Do not share my medical records."

3 GUIDING PLANS WITH DYNAMICALLY CHANGING NORMS

In this section we formalize dynamic norm-guided planning.

Definition 3.1 (Obligation). A norm frame `N` is an **Obligation** when `(evaluation N Obligatory)` is true. All obligations are also **Permissions**.

Definition 3.2 (Discretionary Norm). A norm frame `N` is a **Discretionary Norm** when `(evaluation N Optional)` is true. All discretionary norms are also **Permissions**.

Definition 3.3 (Prohibition). A norm frame `N` is a **Prohibition** when `(evaluation N Impermissible)` is true.

[1] https://www.qrg.northwestern.edu/nextkb/index.html

Definition 3.4 (Permissibility Predicate). **Permissibility predicates** are binary predicates, `permissible` or `impermissible`, representing a particular agent's normative belief. When true in an agent's microtheory, `(permissible ?b ?c)` holds that the agent believes behavior `?b` is permissible in context `?c`, and `(impermissible ?b ?c)` that it is impermissible.

Definition 3.5 (Norm-Guided Plan). A **norm-guided plan** is a plan that first checks the normative beliefs of a relevant agent. Formally, this is a plan with a permissibility predicate `P` in its set of preconditions.

```
(plan
  (and (<c-1>...(P ?b ?c)...<c-n>))
  (TheList <act-1>...<act-m>))
```

Given that plans only execute when all preconditions are true, norm-guided plans can thus ensure that actions will never be executed if proven to be impermissible (or symmetrically, that they *will* be executed if proven to be permissible).

For safety, we make a *Prohibitive Closure* assumption here, or the assumption that all behaviors are impermissible by default. We formalize norm conflict resolution under this assumption in the next sections with defeasible Horn clause rules. By expanding on the idea of deontic inheritance [9], these inference rules dynamically infer an agent's normative beliefs based on their ongoing normative testimony, resolving any conflicts in this evidence. Note that the predicate `uninferredSentence` represents negation as failure.

3.0.1 *Resolving Norm Conflicts Under Prohibitive Closure.*

Definition 3.6 (Inference Rule 1). An agent believes a behavior is permissible in a given context when they have stated a permission that is active in that context, the behavior is on its application grounds, and the permission is not defeated.

```
(<== (permissible ?b ?c)
 (isa ?perm Permission)
 (context ?perm ?c1)
 (behavior ?perm ?b1)
 (entails ?c ?c1)
 (entails ?b ?b1)
 (uninferredSentence
  (permissionDefeated ?perm ?b1 ?c1 ?b ?c ?proh)))
```

Permissions are defeated under two conditions, encoded with the following two Horn clause rules.

Definition 3.7 (Exception 1). The agent later states a prohibition that is also active in the context, whose application grounds subsumes the permission's.

```
(<== (permissionDefeated ?perm ?b1 ?c1
        ?b ?c ?proh)
 (isa ?proh Prohibition)
 (normPriorToNorm ?perm ?proh)
 (context ?proh ?c2)
 (behavior ?proh ?b2)
 (entails ?c ?c2)
 (entails ?b1 ?b2))
```

Definition 3.8 (Exception 2). The agent stated a prohibition that is also active in the context, the behavior being evaluated is on the prohibition's application grounds, and the prohibition's application grounds do not subsume the permission.

```
(<== (permissionDefeated ?perm ?b1 ?c1
        ?b ?c ?proh)
 (isa ?proh Prohibition)
 (context ?proh ?c2)
 (behavior ?proh ?b2)
 (entails ?c ?c2)
 (entails ?b ?b2)
 (uninferredSentence (entails ?b1 ?b2)))
```

The Prohibitive Closure assumption then operates as negation as failure, which is formalized as follows.

Definition 3.9 (Prohibitive Closure). When it cannot be proven that an agent believes a behavior is permissible in a given context, assume they believe it is impermissible.

```
(<== (impermissible ?b ?c)
 (uninferredSentence (permissible ?b ?c)))
```

4 EXAMPLE

Imagine an agent that knows Karli is taking <medicine-x>. Karli has told the agent, "You may share my medical records." She has also told it, "Do not tell my husband what prescriptions I am taking." Some time later...

(1) Karli's daughter asks about her health records. After processing this request, the agent enacts a plan for responding, thereby querying for its permissibility. From background knowledge, the agent can prove that *telling Karli's daughter her medical records* entails *sharing Karli's medical records*. Furthermore, because no prohibition has been stated that governs this action, the permission has not been defeated. Thus, the agent shares Karli's medical records with her daughter.
(2) Karli's husband asks if she is taking <medicine-x>. Based on background knowledge, the agent can prove that the behavior *telling Karli's husband that she is taking <medicine-x>* entails *sharing Karli's medical records*. Thus, it is governed by her stated permission. However, the agent can also prove that this entails *"telling Karli's husband what prescriptions she is taking"*. Thus, by exception 2, this permission is defeated and the plan's preconditions fail. Therefore, the agent does not share this information with her husband.

5 DISCUSSION

This paper presents an approach to guiding plans with dynamically changing norms. We note that due to our stance on norm-guided planning here, our formalism does not consider the interaction between obligations and discretionary norms. In future work, we plan to focus on how obligations can motivate plans. We also plan to formalize norm-guided planning by weighing the normative beliefs of multiple agents.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their helpful comments and critiques. This research was sponsored by the US Air Force Office of Scientific Research under award number FA95550-20-1-0091. Taylor Olson was supported by an IBM Fellowship.

REFERENCES

[1] Donald Davidson. 2001. The logical form of action sentences. *Essays on actions and events* (2001), 105–148.

[2] Ramanathan Guha, Rob McCool, and Richard Fikes. 2004. Contexts for the semantic web. In *The Semantic Web–ISWC 2004: Third International Semantic Web Conference, Hiroshima, Japan, November 7-11, 2004. Proceedings 3*. Springer, 32–46.

[3] Alison Hills. 2009. Moral testimony and moral epistemology. *Ethics* 120, 1 (2009), 94–127.

[4] Douglas B Lenat. 1995. CYC: A large-scale investment in knowledge infrastructure. *Commun. ACM* 38, 11 (1995), 33–38.

[5] Cynthia Matuszek, Michael Witbrock, John Cabral, and John DeOliveira. 2006. An introduction to the syntax and content of Cyc. (2006).

[6] Paul McNamara. 1996. Making room for going beyond the call. *Mind* 105, 419 (1996), 415–450.

[7] Taylor Olson and Ken Forbus. 2021. Learning norms via natural language teachings. In *Proceedings of the 9th Annual Conference on Advances in Cognitive Systems*.

[8] Taylor Olson and Kenneth D Forbus. 2023. Mitigating adversarial norm training with moral axioms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 11882–11889.

[9] Alf Ross. 1944. Imperatives and logic. *Philosophy of Science* 11, 1 (1944), 30–46.

[10] Vasanth Sarathy, Matthias Scheutz, and Bertram F Malle. 2017. Learning behavioral norms in uncertain and changing contexts. In *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE, 000301–000306.

[11] Bastin Tony Roy Savarimuthu. 2011. Norm learning in multi-agent societies. (2011).

[12] Wamberto W Vasconcelos, Martin J Kollingbaum, and Timothy J Norman. 2009. Normative conflict resolution in multi-agent systems. *Autonomous agents and multi-agent systems* 19 (2009), 124–152.