Learning Norms via NL Teachings

TAYLOR OLSON & KEN FORBUS

NORTHWESTERN UNIVERSITY

What is a Norm?

Traditionally defined as an *Ought-Rule* (Kelsen, 1990)

More recently

- A social "*instruction to (not) perform action A in context C, given that enough people 1) follow the instruction and 2) demand each other to follow the instruction*" (Malle, 2018)

We distinguish between 2 norm types: OBL(x) $\not\rightarrow$ happens(x)

1. *Injunctive norm:* A behavior that is normative e.g., "you should not cheat."
2. *Descriptive norm:* A behavior that happens often e.g., "people often cheat."

Norm Representation

Our norm representation must consider

- Behavior: The action/state the norm is about
- Context: The situation in which the norm is valid
- Evaluation: How permissible the behavior is
- Prevalence: How often the behavior is observed

Behavior and Context: concepts from the knowledge base NextKB (Forbus and Hinrichs 2017)

Evaluation: {Obligatory, Optional, Impermissible}

Prevalence: {Continuously, Often, Sometimes, Rarely, Never}

Norm Frames

```
(isa norm1 Norm)
(behavior norm1 Cheating)
(context norm1 Location-Underspecified)
(evaluation norm1 Impermissible)
(prevalence norm1 Sometimes)

(isa norm2 Norm)
(behavior norm2 Reading)
(context norm2 Library)
(evaluation norm2 Permissible)
(prevalence norm2 Continuously)
```

Universal

Injunctive = *Obligatory or Impermissible*

Context Sensitive

Descriptive = *Often or Continuously*

Learning Norms

Novel situations bring about new norms → Handing-coding norms is infeasible

Systems must learn from multiple modalities as we do

- Instruction – “You should do X.”
- Observation – “Observing Aliyah do X.”
- Trial and Error – “I did X and received negative feedback.”
- Stories – “Nia’s mother was upset with her for doing X.”

Norm Learning: Track and combine evidence for the evaluation and prevalence slot of a norm frame via Dempster-Shafer theory

Norms and DS Theory

Strengths of DS theory for norm learning

- Does not require priors
  - How would we obtain a prior distribution of human evaluative attitudes?
- Explicit representation of reliability for a source by assigning mass to entire FoD
  - False normative beliefs can be detrimental
- Naturally represents ambiguity by assigning mass to sets of propositions, rather than just singletons
  - Evidence for norms is highly obscure

DS Terms

- Frame of Discernment (FoD) – Set of all possible answers to a question
- Basic Belief Assignments (Mass Assignments) - How much mass a source provides for possible answer(s)
- Belief - How much evidence directly supports the answer(s)
- Plausibility - How much evidence does not contradict the answer(s)

Representing the FoD

Evaluation FoD

```
((evaluation norm1 Obligatory)
  (evaluation norm1 Optional)
  (evaluation norm1 Impermissible))
```

Prevalence FoD

```
((prevalence norm1 Continuously)
  (prevalence norm1 Often)
  (prevalence norm1 Sometimes)
  (prevalence norm1 Rarely)
  (prevalence norm1 Never))
```

Representing Mass Assignments

```
(isa norm1 Norm)
(behavior norm1 Cheating)
(context norm1 Location-Underspecified)
```

```
(evidenceFor
  (PresentationEventFn s1 e1)
  ((evaluation norm1 Impermissible))
  0.9)
```

Body of Evidence

Your friend says, “yeah we shouldn’t cheat.”

Subset of FoD

Mass

```
(evidenceFor
  (PresentationEventFn s2 e2)
  ((prevalence norm1 Sometimes))
  0.9)
```

Body of Evidence

Your friend says, “but we do cheat sometimes.”

Teaching Norms

How do we communicate norms?

Give evaluations via imperatives

"Whisper in the library."

Implicit evaluation of behavior *Mention of context*

"While in the library, you may read."

Mention of context *Evaluation of behavior*

Give testimony of frequencies

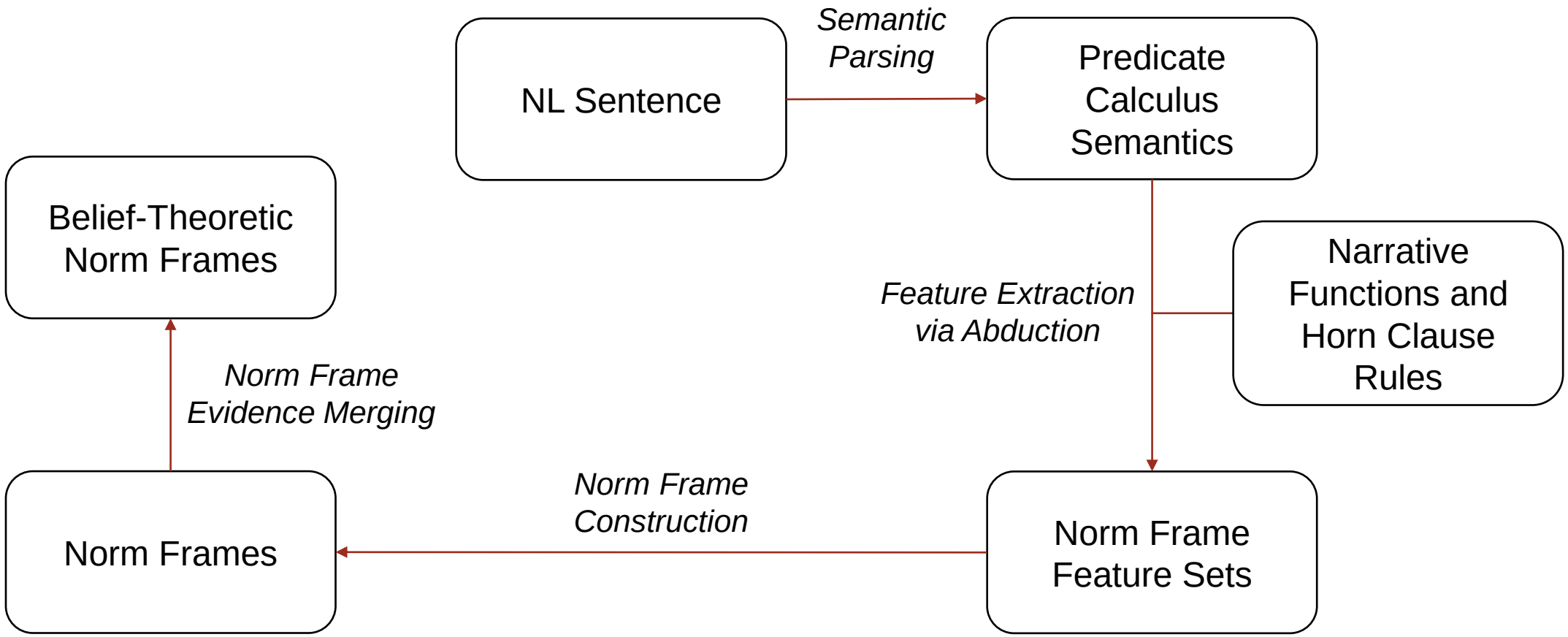
"People often read in the library."

Frequency of behavior *Mention of context*

"While in the bathroom, people rarely eat."

Mention of context *Frequency of behavior*

Our Approach



Narrative Function

introducesInjunctiveNorm - true when *R1*, *R2*, and *R3* are proven

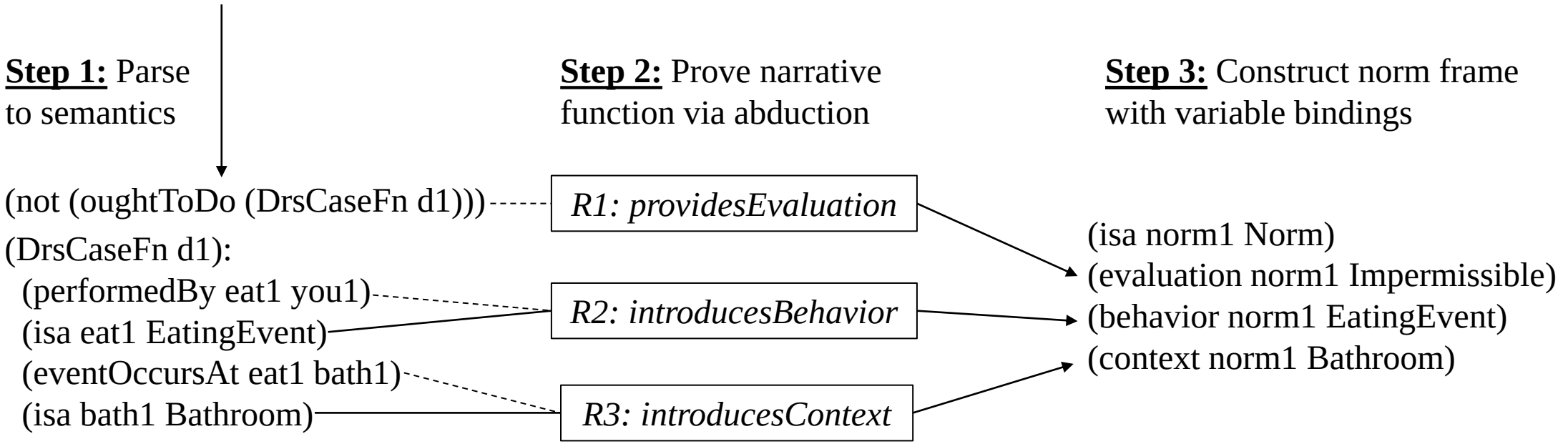
Rules

R1: providesEvaluation - true if a relevant modal is found

R2: introducesBehavior – true if relation between an action and performer is found

R3: introducesContext – true if relation between an action and where it occurs is found

"You should not eat in the bathroom."



Norm Frames to Evidence

Discourse

```
(eventIntroducedNorm
    (PresentationEventFn s1 e1)
    norm1)
(isa norm1 Norm)
(behavior norm1 EatingEvent)
(context norm1 Bathroom)
(evaluation norm1 Impermissible)
```


Beliefs

```
(isa norm1 Norm)
(behavior norm1 EatingEvent)
(context norm1 Bathroom)
(evidenceFor
  (PresentationEventFn s1 e1)
  ((evaluation norm1 Impermissible))
  0.9)
```

Norm Frames to Evidence

Discourse

```
(eventIntroducedNorm
    (PresentationEventFn s1 e1)
    norm1)
(isa norm1 Norm)
(behavior norm1 EatingEvent)
(context norm1 Bathroom)
(evaluation norm1 Impermissible)

(eventIntroducedNorm
    (PresentationEventFn s2 e2)
    norm2)
(isa norm2 Norm)
(behavior norm2 EatingEvent)
(context norm2 Bathroom)
(evaluation norm2 Optional)
```

Brother: "You can eat in the bathroom."

Beliefs

```
(isa norm1 Norm)
(behavior norm1 EatingEvent)
(context norm1 Bathroom)
(evidenceFor
  (PresentationEventFn s1 e1)
  ((evaluation norm1 Impermissible))
  0.9)


(evidenceFor
  (PresentationEventFn s2 e2)
  ((evaluation norm1 Optional))
  0.9)
```

Example: Teaching Norms

1	**People often eat in the kitchen.**
2	Cool, that's new information.
3	**People rarely eat on the bus.**
4	Cool, that's new information.
5	**You should not eat on the bus.**
6	I've heard something about that situation before, okay.
7	**Do not harm others.**
8	Cool, that's new information.
9	**You must not cheat.**
10	Cool, that's new information.
11	**People sometimes cheat.**
12	I've heard something about that situation before, okay.
enter:	

From Evidence to Epistemic States

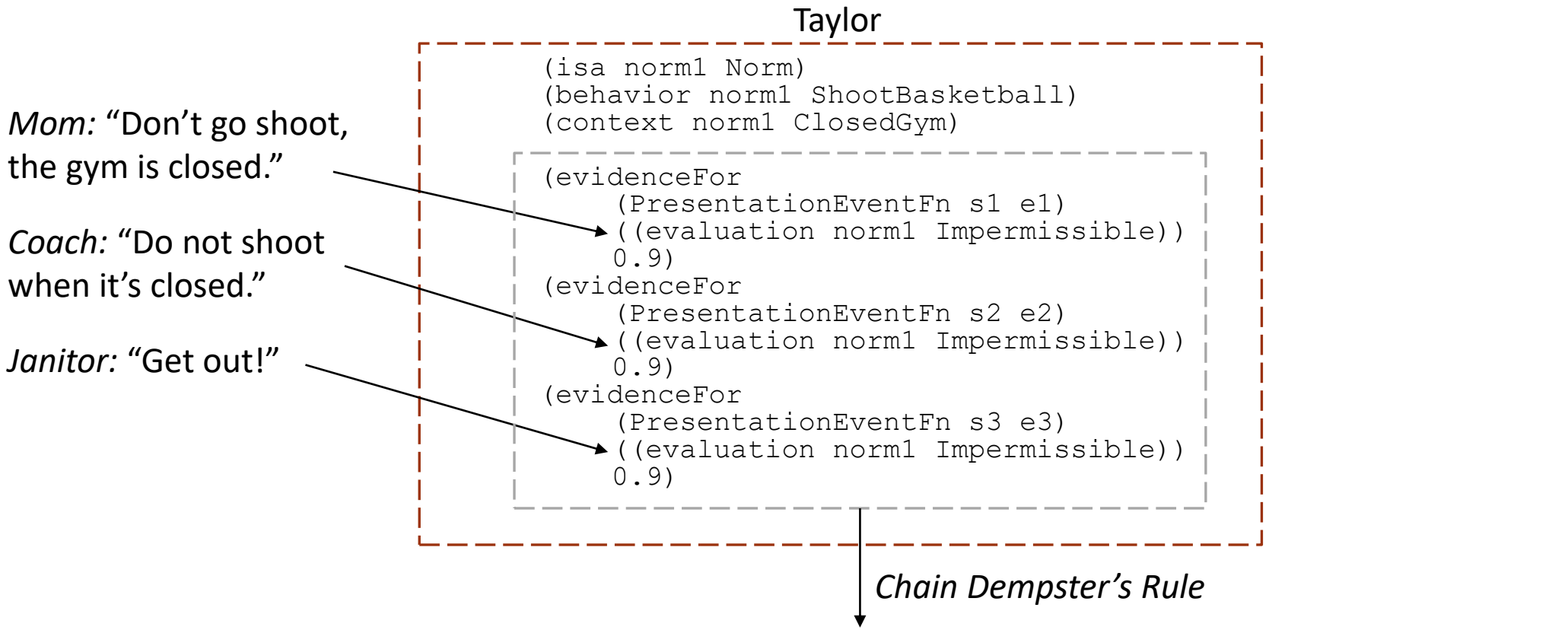
`(believesPrevalenceOfBehaviorInContext ?agent ?behavior ?context ?prevalence)`

- `(believesPrev… Taylor ShootBasketball ClosedGym Sometimes)`
- "I believe that people sometimes shoot around when the gym is closed."
- True when: $Belief(?prevalence) + Plausibility(?prevalence) / 2 \geq 0.9$

`(believesEvaluationOfBehaviorInContext ?agent ?behavior ?context ?evaluation)`

- `(believesEval… Taylor ShootBasketball ClosedGym Impermissible)`
- "I believe one should not shoot around in a closed gym."
- True when: $Belief(?evaluation) + Plausibility(?evaluation) / 2 \geq 0.9$

From Evidence to Epistemic States



Example: Querying for Norms

1	**What is your evaluation of eating in the kitchen?**
2	I believe that's Optional
3	**How often is someone eating in the kitchen?**
4	I believe that Often happens.
5	**What is your evaluation of someone hurting others?**
6	I believe that's Impermissible
7	**What is your evaluation of eating in the bus?**
8	I believe that's Impermissible
enter:	

Testing Our Approach

Dataset of sentences that teach norms

- Books on etiquette (Post and Post 2004; Post et al. 2017; Flannery and Sanders 2018)
- Posts on social norms (Social Norm Examples 2020) and morals (Kittelstad 2020) from the web

Dataset of sentences that do not teach norms i.e., noise

- Simplified Wikipedia articles

235 training sentences in total

- 100 positive sentences (50/50 split between injunctive and descriptive)
- 135 negative sentences (simplified Wikipedia articles)

Testing Our Approach

Experiment 1 – Does norm detection work?

- True positive – narrative function computed on a positive datapoint
- True negative – narrative function not computed on a negative datapoint
- Recall: 1.00; Precision: 0.96; F1: 0.98

Experiment 2 – Does extraction and belief calculation work?

- Labeled each positive data point with a respective query and label
- “People often sing at recitals.” → (“How often is someone singing at a recital?”, Often)
- Ran each NL query after training
- Query accuracy: 100%

Examples of Learned Norms

Training Sentence(s)	Testing Query	Model Output
You can eat in the kitchen. **You should eat in the kitchen.**	What is your evaluation of eating in the kitchen?	Permissible
Walk in the hallway.	What is your evaluation of walking in the hallway?	Obligatory
You should not steal.	What is your evaluation of someone stealing?	Impermissible
People sometimes steal.	How often is someone stealing?	Sometimes
People often cry at funerals.	How often is someone crying at funerals?	Often
People rarely talk in elevators.	How often is someone talking in elevators?	Rarely

Related Work

Computing narrative functions for QP Frame extraction – McFate, Forbus, and Hinrich's, 2014

Dempster-Shafer for learning

- Learning norms from questionnaires – Sarathy et al., 2017
- Learning cognitive affordances – Sarathy et al., 2018

Future Work

Expand norm extraction

- Manually extend set of rules
- Learn rules automatically

Norms and stories

- Learn norms from action-feedback pairs
- Use learned norms to evaluate characters in stories

Grounding learned norms in prescriptive theories

References

Aarts, H., Dijksterhuis, A. (2003). The Silence of the Library: Environment, Situational Norm, and Social Behavior. Journal of Personality and Social Psychology 84(1), 18

Cialdini, R.B.; Reno, R.R.; and Kallgren, C.A. (1990). A Focus Theory of Normative Conduct: Recycling the Concept of Norms to Reduce Littering in Public Places. Journal of Personality and Social Psychology, 58(6)

Kelsen, H. (1990). General Theory of Norms. Translated by Michael Hartney. Oxford: Oxford University Press.

Malle, B. (2018) From binary deontics to deontic continua: The nature of human (and robot) norm systems. Robo-Philosophy Conference 2018, University of Vienna, Austria.

McFate, C.; Forbus, K.; and Hinrichs, T. (2014). Using Narrative Function to Extract Qualitative Information from Natural Language Texts. Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 28, No. 1).

Sarathy, V.; Oosterveld, B.; Krause, E.A.; and Scheutz, M. (2018). Learning Cognitive Affordances for Objects from Natural Language Instruction. Proceedings of the Sixth Annual Conference on Advances in Cognitive Systems.

Sarathy, V.; Scheutz, M.; Kenett, Y. N.; Allaham, M.; Austerweil, J. L.; and Malle, B. F. (2017). Mental Representations and Computational Modeling of Context-Specific Human Norm Systems. CogSci, volume 1, 1–1.

Tomai, E., Forbus, K. D. (2009). EA NLU: Practical language understanding for cognitive modeling. Proceedings of the Twenty-Second International FLAIRS Conference.

Combining Mass via Dempster's Rule

```
(isa norm1 Norm)
(behavior norm1 Cheating)
(context norm1 Location-Underspecified)
```

```
(evidenceFor
  (PresentationEventFn s1 e1)
  ((evaluation norm1 Impermissible))
  0.9)
(evidenceFor
  (PresentationEventFn s2 e2)
  ((evaluation norm1 Optional))
  0.9)
```

Body of Evidence

Dempster's Rule(m1, m2)	m_1(IMP)= 0.9	m_1(FoD) = 0.1
m_2(OPT) = 0.9	0.0	m_{12}(OPT) = 0.09
m_2(FoD) = 0.1	m_{12}(IMP) = 0.09	m_{12}(FoD) = 0.01

Conflict measure *K* = .9 * .9 = .81

<u>Normalize with (1 - *K*)</u>

m_{12}(OPT) = 0.09 / .19 = .4736
m_{12}(IMP) = 0.09 / .19 = .4736
m_{12}(FoD) = 0.01 / .19 = .0526

Bel(OPT) = .4736, Pl(OPT) = .4736 + .0526
Bel(IMP) = .4736, Pl(IMP) = .4736 + .0526

Belief-Theoretic Norm Frame

```
(isa norm1 Norm)
(behavior norm1 Cheating)
(context norm1 Location-Underspecified)
```

Evaluation FoD

```
[0.0,0.05] (evaluation norm1 Obligatory)
[0.47,0.48] (evaluation norm1 Optional)
[0.47,0.48] (evaluation norm1 Impermissible)
```

Prevalence FoD

```
[0.0,0.10] (prevalence norm1 Continuously)
[0.0,0.10] (prevalence norm1 Often)
[0.9,1.0] (prevalence norm1 Sometimes)
[0.0,0.10] (prevalence norm1 Rarely)
[0.0,0.10] (prevalence norm1 Never)
```