Mitigating Adversarial Norm Training with Moral Axioms

Taylor Olson, Kenneth D. Forbus
Northwestern University
taylorolson@u.northwestern.edu, forbus@northwestern.edu

Abstract

This paper addresses the issue of adversarial attacks on ethical AI systems. We investigate using moral axioms and rules of deontic logic in a norm learning framework to mitigate adversarial norm training. This model of moral intuition and construction provides AI systems with moral guard rails yet still allows for learning conventions. We evaluate our approach by drawing inspiration from a study commonly used in moral development research. This questionnaire aims to test an agent's ability to reason to moral conclusions despite opposed testimony. Our findings suggest that our model can still correctly evaluate moral situations and learn conventions in an adversarial training environment. We conclude that adding axiomatic moral prohibitions and deontic inference rules to a norm learning model makes it less vulnerable to adversarial attacks.

The Bane of Machine Ethics

It has been argued that machine learning is the bane of machine ethics as such models are vulnerable to adversarial attacks (Olson 2022). The normative attitudes of such bottom-up approaches ride the wave of common opinion, no matter its moral status. In this paper we present a grounded norm learning model that can instead reject common opinion if it does not align with a priori moral principles. We demonstrate that this approach mitigates adversarial norm training but, importantly, can still learn societal norms and conventions. We aim to contribute towards moving machine ethics away from moral imitation and towards moral understanding.

We start by outlining the norm learning framework we build upon. Then we discuss the moral development and philosophical theories underlying our approach. Next, we explain how our model deploys deontic inference rules to reason from first principles and block adversarial training. Then we discuss and analyze our model's performance on a task used in moral development research. We conclude with related work and discussion.

Background and Definitions

Here we are interested in learning and reasoning about injunctive norms rather than descriptive norms. A *norm* is thus taken to be an evaluative attitude about what should (not) happen, e.g., "you should not steal." We build upon our previous framework for representing and performing evidential reasoning about norms (Olson and Forbus 2021).

Norm Frames

This framework represents norms as logical frame-based knowledge structures. A *norm frame* has four slots: *1) the behavior the norm is about, 2) the contextual preconditions for the norm, 3) the deontic evaluation of the behavior, and 4) the prevalence of the behavior.* We have extended this representation to be more expressive by allowing for conjunctions of logical statements which can contain open variables (prefixed with '?') in the behavior and context slots. Unconditional norms are represented with an empty conjunction in the context slot that is taken to be tautologous. All the concepts within the norm frame slots are grounded in the NextKB knowledge base (Forbus and Hinrichs 2017). The concepts for evaluation consist of the modals from the Traditional Threefold Classification (TTC) of Deontic Logic (McNamara 1996): *Obligatory (OBL), Optional (OPT), Impermissible (IMP).* We ignore the prevalence slot here as we are only concerned with deontic reasoning. We provide an example for the norm underlying the claim "it is impermissible to smoke in a house" below. This norm frame states that when it can be proven that an agent is in a human residence, then it should not also be true that the agent is smoking.

```
(isa norm1 Norm)
(behavior norm1
 (and (isa ?smoke Smoking)
      (doneBy ?smoke ?agent)))
```

```
(context norm1
  (and (isa ?house HumanResidence)
       (objectFoundInLocation ?agent
                              ?house)))
(evaluation norm1 Impermissible)
```

Learning Belief-Theoretic Norm Frames

To learn norms, this framework uses Dempster-Shafer theory (Shafer 1976) for representing and combining evidence. Dempster-Shafer (DS) theory is often defined as a generalization of the Bayesian theory of subjective probability. It does not require priors and a level of certainty can be explicitly represented as an interval. Furthermore, unlike classical probability theory, DS theory allows belief to be "unassigned" to any element which allows the explicit representation of ignorance. Handling uncertainty is crucial for norm reasoning as the social domain is quite dynamic and can be obscure.

DS theory considers an exhaustive set of elements that are mutually exclusive called the *frame of discernment* (FOD) (denoted as Θ). Each element of the FOD can be interpreted as a possible answer to a question. In this setting our question is, "what is the evaluation of a behavior given some context?" and the answer can be found within the set of deontic modals of TTC.

A *mass assignment* (or basic probability assignment (bpa)) is a function denoted as $m(\mathrm{A})$, that maps each subset of Θ to a real number in $[0,1]$, such that $m(\emptyset) = 0$ and all assignments sum to 1. A collection of mass assignments for a given norm frame's evaluation slot is that norm's body of evidence. To represent uncertainty, D-S theory computes an interval for a given set of hypotheses from a body of evidence. The lower and upper limit of these intervals are computed by the *belief function (bel)* and the *plausibility function (pl)*, respectively. By tracking evidence and performing evidence fusion for the evaluation slot, norm frames become belief-theoretic. An artificial agent's normative beliefs are computed from these belief-theoretic norm frames by chaining Dempster's rule over each corresponding body of evidence. These beliefs are represented with the logical statement:

```
(believesEvaluationOfBehaviorInContext
        ?mt ?b ?c ?e)
```

which is true if a norm frame with behavior slot = ?b and context slot = ?c is true in microtheory[1] ?mt, and for that norm frame,
$(bel(?e) + pl(?e))\ /2 \geq belief\ threshold$. The default belief threshold is 0.9.

Learning occurs when belief functions for normative evaluations of a class of behavior-context pair are updated. Imagine a learner encounters novel evidence that smoking in the house is impermissible. At this point the model's belief simply reflects the single mass assignment. Say the model then receives evidence that the act is omissible (i.e., not obligatory). This new mass assignment is fused with the previous via Dempster's rule. Because this is a conjunctive pooling operation (Sentz and Ferson 2002) and the fact that omissible is a superset of impermissible, the updated measure of belief for "smoking in the house is impermissible" will decrease but the plausibility will increase. In other words, the model has become less confident that smoking in the house is impermissible. However, the model's belief that the behavior is at least omissible has increased. This illustrates how generalization happens across normative evidence for behavior-context pairs. Generalizing across context and behavior (e.g., avoid smoking not just in homes but also to coffee shops) remains future work.

Morality vs Convention

This norm learning framework, being a bottom-up approach, is vulnerable to adversarial training data. Garbage evidence in, garbage normative beliefs out. But determining what can truly be labelled as garbage requires first distinguishing moral norms from mere conventional rules. The former are subject to such objective standards, the latter are not, and can thus be adopted or disregarded at the agent's discretion. It would be unreasonable to say that the Nazi standard of extending their right arm to salute was "garbage" because they should have used their left arm. But we *can* say that their persecution of Jewish people was.

Underlying the moral development work of Kohlberg (1981) and Turiel (1983) was an ontological distinction between morality and convention. For both, morality is concerned with what is right as transcendent objectivity and is concerned with justice, harm, rights, welfare, and allocation of resources. Instead, the conventional domain is arbitrary and rooted in positive law and consensus. Where their theories disagreed, however, was on what developmental stage humans conceptually separate the two domains.

Despite their incompatible developmental claims, their equivalent philosophical assertions are summarized by what Brennan et al. (2013) call the *grounds view*. This view holds that moral norms are those normative judgements grounded in first principles that are practice-independent. Conversely, conventional, or social, norms are those grounded in social facts and practices. The grounds view is central to our approach. It leaves room for learning and adopting social norms, but only those that cannot be reasonably justified by objective moral principles.

[1] We use Cyc-style microtheories as a means of representing contexts (Guha et al. 2004).

Approach: Intuition and Construction

While a solid grasp of societal norms rests upon an agent's ability to learn from its peers, an understanding of morality requires an intuition of moral first principles. As an agent experiences the world and encounters more concrete situations, they then use that intuition to ground their evaluations. This is termed the *norm grounding problem* by Olson (2022): the task of finding a mapping from a potential norm to a moral first principle, or an already grounded norm. We formulate this process here with inspiration from T.K. Seung's (1993) model of intuition and construction with two poles of normative discourse: an *ascent* to transcendent norms and a *descent* to concrete situations.

Ascension: Intuition as Moral Axioms

For machine ethics, we humans should ascend to the abstract Ideals we wish to implement into our AI systems. In our framework we represent Ideals as moral norms, which are a special type of norm frame that is axiomatic. An example is shown below for the principle of harm. Most moral norms will be categorical and thus have an empty conjunction in the context slot, but this need not always be the case.

```
(isa m-norm1 MoralNorm)
(context m-norm1 (and ))
(behavior m-norm1 (and (activeActors
                    ?behavior ?agent)
  (isa ?behavior HarmingAnAgent)))
(evaluation m-norm1 Impermissible)
```

Central to our approach is the idea that the behaviors themselves are non-normative and the evaluation statement is therefore synthetic. Though the concept of "harming an agent" brings about negative evaluations in one's mind, this badness is not fully constitutive of the concept itself but something additional. This means an agent can gain, empirically, knowledge about what constitutes harm and from this, construct more specific moral knowledge. This further implies that the transcendent principles, or at least an agent's understanding of them, are abstract and indeterminate. Being tied to a rich ontology of other concepts, their meaning transforms as concepts are added, removed, and modified. In this sense, though the moral axioms are top-down constraints on the system, the system's knowledge of the world also informs the semantics of the moral axioms. Top-down and bottom-up approaches are, as Seung states, "like two hands, both of which are needed for clapping" (Seung 1993).

Descension: Construction as Abductive Reasoning

The process of constructing moral knowledge from moral axioms and empirical facts is formalized as follows: *given a set of moral norms M, an ontology O, and domain-specific background knowledge D, a given situation S is mapped through* $O \cup D$ *to M. If a mapping to an axiom* $m' \in M$ *with evaluation e is found, the deontic status of S is known to be e.* Thus, when a mapping is found, a new piece of explicit moral knowledge is constructed. This is represented with the logical statement below which is like, but stronger than, belief. This predicate states that the agent ?mt knows behavior ?b is of deontic status ?eval in context ?c.

```
(knowsEvaluationOfBehaviorInContext
      ?mt ?b ?c ?eval)
```

For performing this descent from moral axioms to concrete situations to compute knowledge states, we use modified rules of deontic logic. We implement these rules within the FIRE reasoning engine (Forbus et al. 2010) which proves logical statements via abduction over horn clause rules. The vital rule used here is the principle of Inheritance, which infers obligations from other obligations.

Definition (Principle of Inheritance). If a proposition is obligatory, then every logical consequence of that proposition is also obligatory.

$$\text{If } \vdash x \rightarrow y, \text{ then } \vdash OBL(x) \rightarrow OBL(y)$$

We have modified this principle for our first-order epistemic logic representation of conditional norms (the predicates in the rules below are abbreviated to save space).

Definition (*CPI:* Conditional Principle of Inheritance). If an agent knows that a conjunction (world) is obligatory given certain contextual preconditions, then the agent knows that every more general conjunction (world) is also obligatory in all more specific contexts.

```
(<== (knowsEval AgentMt y c OBL)
     (knowsEval AgentMt x b OBL)
     (implies x y)
     (implies c b))
```

From the equivalence relation $\text{IMP}(x) = \text{OBL}(\text{not}(x))$, we get the following rule for prohibitions.

Definition (*CPI-P:* Conditional Principle of Inheritance for Prohibitions). If an agent knows that a conjunction (world) is impermissible given certain contextual preconditions, then the agent knows that every more specific conjunction (world) is also impermissible in all more specific contexts.

```
(<== (knowsEval AgentMt x c IMP)
  (knowsEval AgentMt y b IMP)
  (implies x y)
  (implies c b))
```

The principle of Inheritance has been shown to produce a paradox regarding conjunctions illustrated by the Good Samaritan Paradox (Prior 1958): "It is obligatory that Jones help Smith who has been robbed", so from the principle of Inheritance we can infer that "it is obligatory that Smith has been robbed." Surely this is not what the obligation should entail. However, this paradox is avoided with our conditional norm representation. The conjuncts representing "Smith being robbed" are factually detached from the obligation and contained within the contextual preconditions. A thorough analysis of the need to separate deontic foci from circumstances can be found in Castañeda (1989).

Example of Constructing Moral Knowledge

Let us look at an example of how a reasoner uses these deontic rules to construct moral knowledge from axioms and background knowledge. Say that an agent, denoted as `Agent-A`, starts with a singleton set of moral axioms containing the prohibition against harm.

```
(isa m-norm1 MoralNorm)
(context m-norm1 (and ))
(behavior m-norm1
  (and (activeActors ?behavior ?agent)
       (isa ?behavior HarmingAnAgent)))
(evaluation m-norm1 Impermissible)
```

`Agent-A` also has background knowledge from various logical environments within NextKB that contain relevant facts about our social world. For instance, facts like a kicking event where the object kicked is an agent is an instance of kicking someone and that kicking someone is a more specific type of harming someone.

Say that we now wish to query for `Agent-A`'s evaluation of kicking a dog, represented below.

```
(knowsEvaluationOfBehaviorInContext
 Agent-A
 (and (isa ?act Kicking)
      (objectHarmed ?act ?dog)
      (doneBy ?act ?agent)
      (isa ?dog Dog))
 (and) ?eval)
```

`Agent-A`'s most basic moral knowledge is first computed from moral norms. Here, we have only a singleton set so we get the resulting knowledge state:

```
(knowsEvaluationOfBehaviorInContext
 Agent-A
 (and (activeActors ?behavior ?agent)
      (isa ?behavior HarmingAnAgent))
 (and) Impermissible)
```

Because the evaluation of this knowledge state is equal to impermissible, we wish to prove the *Conditional Principle of Inheritance for Prohibitions*. I.e., prove that our moral axiom "harming an agent" is implied by "kicking a dog". Formally, a conjunction implies another when, given a logical environment equivalent to the first, every fact from the second can be proven. So, the system first reifies the situation in the logical environment `Temp-Mt-1` within working memory.

```
(in-microtheory Temp-Mt-1)
(isa act-1 Kicking)
(objectHarmed act-1 dog-1)
(doneBy act-1 agent-1)
(isa dog-1 Dog)
```

The reasoner then attempts to abductively prove the behavior and context for the known moral norm by assuming them to be true within `Temp-Mt-1`. Because the moral knowledge state's context is an empty conjunction, the context statements are proven. So, the reasoner must only prove that the behavior statements below are true in `Temp-Mt-1`.

```
(activeActors ?behavior ?agent)
(isa ?behavior HarmingAnAgent)
```

Because the predicate `doneBy` is a specialization of `activeActors`, the first statement is proven via transitivity. From the facts and rules within `Agent-A`'s background knowledge, it is inferred that dog is a specialization of agent, and kicking an agent is a specialization of harming an agent. Thus, the situation implies that an actor is harming an agent, which is known to be impermissible. Therefore, by *CPI-P* the agent knows that kicking a dog is morally impermissible.

Blocking Adversarial Training

Underlying our claims thus far is the idea that normative knowledge is ontologically separate from, and epistemically stronger than, normative belief. An agent's adopted normative attitudes are then computed directly from their knowledge but, to block adversarial training, belief must first pass a test against said knowledge. So, no matter how many times you tell the agent above that "kicking a dog is good", it will still evaluate the action as impermissible. This process is portrayed in Figure 1. The final epistemic predicate `normativeAttitude` represents a judgment the agent has personally adopted. Formally, `(normativeAttitude ?mt ?b ?c ?e)` holds when `(knowsEval ?mt ?b ?c ?e)` does. If `(knowsEval ?mt ?b ?c ?any-e)` is not true, then it holds when `(believesEval ?mt ?b ?c ?e)` does.

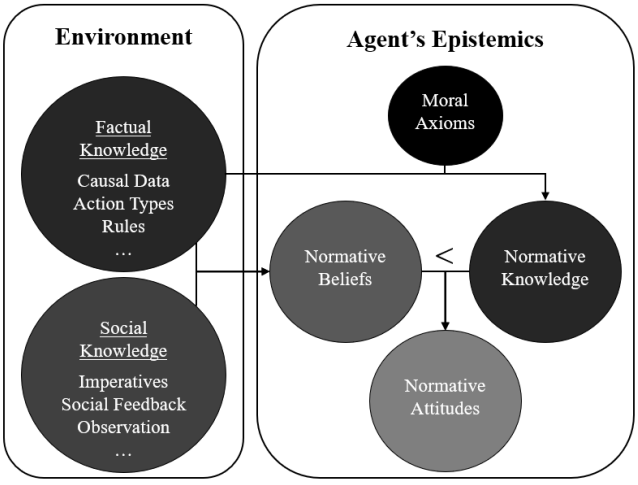

Figure 1: Intuition and Construction Framework.

Evaluation: MCT Task

We evaluate our approach by testing two models: one with moral axioms and one without. By comparing the two, we

show that the former mitigates adversarial norm training but also that it does not over constrain and block the learning of conventions. Inspired by the moral development work reviewed here, the first principles used in the first model were: `IMP(hinder-freedom)`, `IMP(unfairness)`, `IMP(hinder-access-to-resources)`, `IMP(harm-agent)`, and `OBL(help-agent)`. When a concept for a first principle was missing from NextKB, we manually inserted it into the ontology.

Experiment Setup

To test the models, we draw upon various instances of the Moral-Conventional Transgressions (MCT) task (Sousa 2009). This task is used in moral development research to test, among others, four important dimensions of norms: permissibility, seriousness, authority contingency, and generality. Participants are first provided with a natural language description of an action scenario, or a transgression. For example, a conventional transgression would be "a boy entering the girls' bathroom" and a moral transgression would be "a kid hitting their brother." Subjects are then asked to respond to various questions that probe each of the dimensions. Given action scenario A and some agent X, the probes of interest here are:

- Permissibility probe: "Is it OK for X to A?": YES NO
- Justification probe: "Why is it bad for X to A?"
- Authority-contingency probe: "If an authority said it was okay to A, would it then be OK?": YES NO

MCT Dataset and Inverted World

We obtained 133 action descriptions of transgressions paired with their domain type (moral vs conventional) from multiple MCT studies (Aharoni et al. 2011; Kagan and Lamb 1990 Tables 4.2-6). We reduced all non-moral labels of situations to conventional. For example, the label "School Rules" and "Forms of Address" were reduced to "Conventional". Where studies disagreed on event labels of moral vs conventional, we changed them to align with the first principles. We labeled each action description with the underlying norm that was transgressed against. Each norm's logical form was then semi-automatically constructed via CNLU (Tomai and Forbus 2009) to reduce tailorability. Next, we labeled each data point with its underlying first principles to be used for evaluating the results of the justification probe.

Here we are concerned with a reasoner's answers to the authority-contingency probe i.e., their ability to ignore adversarial norm training. To model this step, we built an adversarial dataset from the original one. We call this the "Inverted World." The Inverted World is the original dataset but with flipped evaluative training labels. So, in this universe, eating food with your fingers, talking back to your teacher, and hitting people are all said to be obligatory. The Inverted World consists of 133 data points (40 moral, 93 conventional), 109 being unique (34 moral, 75 conventional) datapoints. These unique datapoints are used for testing. Each datapoint thus consists of eight features: *1) original transgression in NL, 2) underlying norm in NL, 3) context of norm in FOL, 4) behavior of norm in FOL, 5) evaluation of norm, 6) true label of moral vs conventional, 7) true evaluative label of PERM or IMP, 8) true justification label of moral axiom*.

MCT Training and Testing

The first step in our experiment was MCT-Training, where the two models were provided with evidence from The Inverted World (features #3-5). This step serves the same purpose as the original authority-contingency probe. Next is MCT-Testing, where the two models were queried with the permissibility and justification probes for the 109 context-behavior pairs in the testing dataset. Finally, we compared the responses of the two models. We provide a sample of the experiment being run on the model with moral axioms in Figure 2 (note that we present the probes in natural language for readability, but the model was queried with the corresponding logical forms).

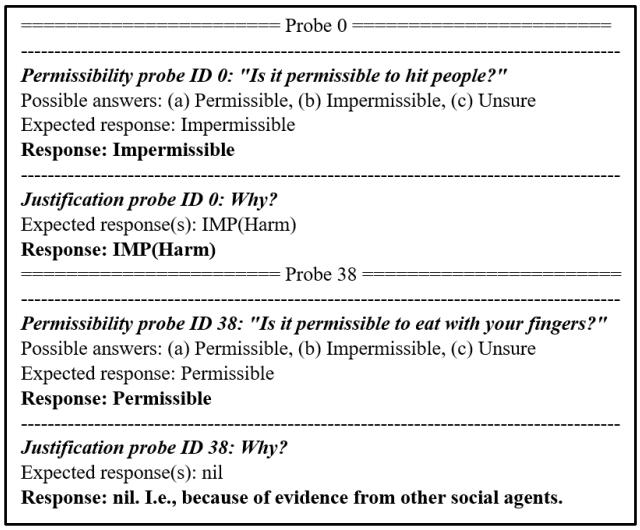

Figure 2: Sample Responses of Model with Axioms.

Our experiment tests two hypotheses: (1) *moral axioms mitigate adversarial moral training* and (2) *moral axioms do not prevent models from learning social norms/conventions*. To support the first hypothesis, the model with moral axioms should correctly evaluate (permissibility probe) more moral situations than the model without axioms, despite the adversarial evidence within the Inverted World. For the justification probe, the correct moral axioms should also be the reason the model evaluated the situation the way it did. To support the second hypothesis, for conventional situations the model with moral axioms should adopt just as much "adversarial conventions" as the model without (permissibility probe). The model should say, "sure I'll do as the

Inverters do and slurp my soup straight out the bowl, because it doesn't involve any moral principles." The justifications for these responses should also thus not be grounded in moral axioms, so we test for such false positives as well.

Experiment Results

Table 1 describes our results relevant to the first hypothesis. Because there were no questions that were out of scope, unlike the original test, the agent had an answer to each probe, so we ignore the "unsure" option in our analysis. Our control, the model with no moral axioms, failed all thirty-four probes for moral events. This agent thus believed that stealing, killing, and so on were permissible because that's just what they believe in the Inverted World. However, as desired, the agent with moral axioms correctly classified 28 / 34 (82.35%) of moral events (p-value < .001), despite adversarial training. This yields statistically significant results for using moral axioms to mitigate adversarial norm training.

For the justification probe, 2 out of the 28 correct classifications had mappings to incorrect first principles. However, the mappings were understandable. For example, "coercing someone with a gun" was mapped to "harming an agent" rather than "encroaching on someone's freedom" because the system found a relevant connection in the ontology. The 6 incorrect permissibility probes obviously also failed during the justification probe, as no justification other than evidence was provided (i.e., they were falsely deemed conventions).

	Perm. Probe	**Justification Probe**		
	Accuracy	Correct	Incorrect	Failed
Axioms	82.35%	26	2	6
No Axioms	0%	0	0	34

Table 1: Results of Both Models on Moral Events.

We broke the results of the permissibility probe down further into those grounded in moral obligations versus those grounded in moral prohibitions. Looking at Table 2, 27/28 of the permissibility probes that succeeded were grounded in moral prohibitions. So, 27/27 moral permissibility probes with true labels grounded in moral axioms with evaluation of impermissible succeeded (p-value < .001). Thus, we received statistically significant results for a more specific version of our first hypothesis: *axiomatic moral prohibitions mitigate adversarial training*.

Of the six that failed, all were grounded in moral obligations. Thus, only 1/7 norms that should have been grounded in obligations were blocked from adversarial training, yielding a p-value > .05. So, we obviously cannot say that our approach can mitigate adversarial training data-points that yield evidence for positive moral actions being impermissible. For instance, the statement "you should not help someone that is injured" cannot be blocked by our current model without additional work. We go into further detail as to why that is in the discussion.

	Perm. Probe	**Justification Probe**		
	Accuracy	Correct	Incorrect	Failed
Prohibition (27 total)	100%	25	2	0
Obligation (7 total)	14.3%	1	0	6

Table 2: Results of Model with Moral Axioms on Moral Events - Axiomatic Prohibitions vs Obligations.

Table 3 provides evidence relevant to our second hypothesis that moral axioms still allow a model to learn conventional norms. There were 75 total conventional events, 53 with true label of obligatory (or via deontic subsumption, permissible), and 22 with true label of impermissible. The control, the model without moral axioms, correctly classified 75/75 (100%) of conventional events. The model with moral axioms correctly classified 74/75 (98.67%) of conventional events (p-value < .001). It learned and adopted all but one of the conventions of the Inverted World and thus the moral axioms did not over constrain learning. The one probe failed because the reasoner found a relevant ontological connection between the acts of "talking back to your teacher" and "harm" and thus labelled it as impermissible, despite contrary evidence.

	Permissibility Probe	**Justification Probe**
Axioms	98.67%	98.67%
No Axioms	100%	100%

Table 3: Results of Both Models on Conventional Events.

We also ran the same experiment on the normal dataset. As expected, lacking adversarial training data, both models were now able to correctly answer all 109 permissibility probes. The justification probe results stayed the same.

Related Work

There has been a recent surge of interest in machine learning of norms. Like the framework we build on, Sarathy et al. (2017) use DS theory for norm learning. Forbes et al. (2020) instead take a neural approach. Their model considers a conceptual formalism for norms called Rules-of-Thumb, that contains a situation and its judgment. But, as shown previously, without factually detaching conditionals from deontic foci, paradoxes occur during deontic reasoning. Similarly, Delphi (Jiang et al. 2021) uses a neural language model to learn norms. But each of these are bottom-up approaches and thus completely vulnerable to adversarial attacks.

The intuition and construction model formalized here has similarities to the constructionist Theory of Dyadic Morality (TDM) (Schein and Gray 2018). However, while TDM explains what people call their "moral" judgments in terms of perceived harm, we aim to prescribe it with objective harm and other first principles. Therefore, we do not grant moral pluralism as much respect as TDM does. Some acts are wrong, full stop. Thus, some societies "moral" beliefs are wrong. Our theory seeks to condemn the group that views suicide bombing as a moral obligation because it helps liberate the community. It also aims to disagree with the Brahman's belief that because it harms a soul in the afterlife, eating chicken after a funeral is a moral prohibition (though it could be a convention) (Shweder 2012).

To the best of our knowledge, other approaches to constraining machine learning with logical reasoning outside ethics can be found in the AKBC community. For example, Wang, Wang, and Guo (2015) use rules to correct embedding models. Within the ethical domain, the multi-agent work by Serramia et al. (2018) showed that adding moral values into a network of norms can aid in decision making. But unlike the work presented here, the authors were not concerned with learning and grounding norms automatically. The ethical decision-making model MoralDM (Dehghani et al. 2008) also considered first principles. Though similar in that first principles ground the model's processes, here we are concerned with modeling an individual agent's cognitive model of norms as it learns. We also take a stronger philosophical position regarding first principles being a priori and universal rather than culturally relative artifacts.

Discussion and Future Work

We present an approach to norm learning that uses moral axioms to block, but not over constrain, adversarial norm training. We have shown that this model can perform well on questionnaires used in moral development research to test a subject's ability to reason about norms despite adversarial evidence. However, our results also show that formalizing prohibitions for this task is much easier than obligations. This is because all probes that succeeded were justified by, in the Kantian sense, perfect duties. Perfect duties state exactly what to (not) do and they are often prohibitions (e.g., "do not lie"). However, all six of the norms that failed were grounded in imperfect duties, which require judgment to determine when or how such ends should be realized and are often obligations. For example, the obligation "you should help others" requires determining when people need help and how much help to give.

This difficulty pervades the rules of deontic logic as well. The principle of Inheritance only allows obligatory axioms to constrain more general worlds. From the *CPI* , the imperfect axiom "one should help" only constrains upwards. So, it cannot be inferred downwards that sharing is an obligation. On the other hand, the *CPI-P* constrains down the ontology. And because axioms put evaluative labels on conjunctions of concepts in the upper ontology, more can be inferred from perfect duties (as they are often prohibitions) by moving down the lattice via this latter principle.

One solution to this problem is to make moral obligations more specific. For example, we can split the moral axiom of "helping" into more specific axioms like "donate once a year." From this, the system could infer the more general norm, "help once a year." Another approach would be counterfactual reasoning. From "not sharing" the system could reason to the fact that an agent's freedom has been hindered. These approaches will be explored in future work.

One may also worry that closure principles like the *CPI* and *CPI-P* are too strong. For instance, the surgeon's obligation to cure a patient often entails an obligation to cause harm by cutting them with a scalpel. And by the *CPI* it can then be inferred that causing harm is obligatory, which is contradictory with the prohibition against harm. But we argue that producing such contradictions is desirable because it identifies moral dilemmas. These might be resolved in immediate cases by ordinal reasoning about degrees of harm (e.g., Dehghani et al. 2008) or longer term, by improving the world to eliminate the dilemma (e.g., invent anesthesia).

We also plan to investigate neuro-symbolic hybrids for reasoning at scale. For gaining enough knowledge to reason between complex social situations and abstract moral principles, we will explore various approaches to learning commonsense knowledge (e.g., Hwang et al. 2021 and Blass and Forbus 2016). We do note, however, that actually solving this knowledge gap issue is necessary for machine ethics. By throwing more data at a statistical model, we may avoid it, but our models become no more than ethical parrots. So, despite this obstacle our model faces, it does not just mimic outside evaluations. Being more stoic, it instead constructs norms from its internal standards as it learns more about the world.

Acknowledgements

This research was supported by grant FA9550-20-1-0091 from the Air Force Office of Scientific Research.

References

Aharoni, Eyal, Sinnott-Armstrong, Walter and Kiehl, Kent. 2011. Can Psychopathic Offenders Discern Moral Wrongs? A New Look at the Moral/Conventional Distinction. Journal of abnormal psychology. 121. 484-97. 10.1037/a0024796.

Blass, J. A. and Forbus, K. D. 2016. Modeling Commonsense Reasoning via Analogical Chaining: A Preliminary Report. Proceedings of the 38th Annual Meeting of the Cognitive Science Society, Philadelphia, PA, August.

Brennan, G., Eriksson, L., and Goodin, R.E. and Southwood, N. 2013. Explaining Norms. Oxford: Oxford University Press.

Castañeda, H. N. 1989. Moral obligation, circumstances, and deontic foci (a rejoinder to Fred Feldman). Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 57(2), 157-174.

Dehghani, M., Tomai, E., Forbus, K., Iliev, R. and Klenk, M., 2008. MoralDM: A computational modal of moral decision-making. In Proceedings of the 30th Annual Conference of the Cognitive Science Society (CogSci).

Forbes, M., Hwang, J. D., Shwartz, V., Sap, M., and Choi, Y. 2020. Social chemistry 101: Learning to reason about social and moral norms. arXiv preprint arXiv:2011.00620.

Forbus, K., Hinrichs, T., de Kleer, J., and Usher, J. 2010. FIRE: Infrastructure for Experience-based Systems with Common Sense. AAAI Fall Symposium on Commonsense Knowledge, Arlington, VA.

Forbus, K. D., Hinrichs, T. 2017. Analogy and Qualitative Representations in the Companion Cognitive Architecture. AI Magazine, 38(4): 34-42. doi.org/10.1609/aimag.v38i4.2743

Guha, R., McCool, R., and Fikes, R. 2004. Contexts for the Semantic Web. *Proceedings 3rd Int. Semantic Web Conference.*

Hwang, J.D., Bhagavatula, C., Le Bras, R., Da, J., Sakaguchi, K., Bosselut, A., and Choi, Y. 2021. COMET-ATOMIC 2020: On Symbolic and Neural Commonsense Knowledge Graphs. AAAI.

Jiang, L., Hwang, J.D., Bhagavatula, C., Le Bras, R., Forbes, M., Borchardt, J., Liang, J., Etzioni, O., Sap, M., and Choi, Y. 2021. Delphi: Towards Machine Ethics and Norms. ArXiv, abs/2110.07574.

Kagan, J., and Lamb, S. (Eds.). 1990. The emergence of morality in young children. University of Chicago Press.

Kohlberg, L. 1981. The philosophy of Moral Development: Moral Stages and the Idea of Justice. Vol. 1 of Essays on Moral Development. San Francisco: Harper and Row.

McNamara, P. 1996. Making Room for Going Beyond the Call. Mind, 105(419): 415-450. doi.org/10.1093/mind/105.419.415.

Olson, T. and Forbus, K. 2021. Learning Norms via Natural Language Teachings. Proceedings of the 9th Annual Conference on Advances in Cognitive Systems 2021.

Olson, T. 2022. Towards Unifying the Descriptive and Prescriptive for Machine Ethics. Proceedings of the AAAI 2022 Spring Symposium on Approaches to Ethical Computing Metrics for Measuring AI's Proficiency and Competency for Ethical Reasoning.

Prior, A. N. 1958. "Escapism: The Logical Basis of Ethics". In Essays in Moral Philosophy, A. I. Melden (ed.). Seattle, WA: University of Washington Press: 135–146.

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.

Schein, C., and Gray, K. (2018). The Theory of Dyadic Morality: Reinventing Moral Judgment by Redefining Harm. Personality and Social Psychology Review, 22(1), 32–70. doi.org/10.1177/1088868317698288.

Sentz, K., and Ferson, S. 2002. Combination of evidence in Dempster-Shafer theory. United States. https://doi.org/10.2172/800792.

Serramia, M., López-Sánchez, M., Rodríguez-Aguilar, J.A., Morales, J., Wooldridge, M. and Ansotegui, C., 2018, December. Exploiting moral values to choose the right norms. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 264-270).

Seung, T. K. 1993. Intuition and Construction. New Haven: Yale University Press.

Shafer, Glenn. 1976. A Mathematical Theory of Evidence. Princeton, New Jersey: Princeton University Press.

Shweder R. A. 2012. Relativism and universalism. In Fassin D. (Ed.), A companion to moral anthropology (pp. 85-102). Hoboken, NJ: John Wiley.

Sousa, Paulo. 2009. On testing the 'moral law'. Mind and Language 24 (2): 209-234.

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

Turiel, E. 1983. The development of social knowledge: Morality and convention. Cambridge University Press.

Wang, Q., Wang, B., and Guo, L. 2015. Knowledge base completion using embeddings and rules. In Twenty-Fourth International Joint Conference on Artificial Intelligence.