Normative Testimony and Belief Functions: A Formal Theory of Norm Learning

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

Abstract

The ability to learn another's moral beliefs is necessary for all social agents. It allows us to predict their behavior and is a prerequisite to correcting their beliefs if they are incorrect. To make AI systems more socially competent, a formal theory for learning internal normative beliefs is thus needed. However, to the best of our knowledge, a philosophically justified formal theory for this process does not yet exist. This paper begins the development of such a theory, focusing on learning from testimony. We make four main contributions. First, we provide a set of axioms that any such theory must satisfy. Second, we provide justification for belief functions, as opposed to traditional probability theory, for modeling norm learning. Third, we construct a novel learning function that satisfies these axioms. Fourth, we provide a complexity analysis of this formalism and proof that deontic rules are sound under its semantics. This paper thus serves as a theoretical contribution towards modeling learning norms from testimony, paving the road towards more social AI systems.

1 Introduction

It is widely accepted that our AI systems need to know what we humans value to effectively participate in our society. Research on modeling such evaluative phenomena is relatively new, but helpful analogies have been made to philosophical research in descriptive ethics [Jiang, L. et al., 2021; Olson, T., 2024]. The descriptive ethicist asks, "what does this population believe is good/bad/neutral?" To answer this, they gather relevant evidence and infer the population's normative beliefs. We work within this analogy here to construct a theoretically grounded formalism for norm learning.

We specifically consider learning from normative testimony, i.e., natural language expressions such as "you should help others". Independent of whether one is an optimist or pessimist about normative testimony [Hills, A., 2009] , these acts relay the norms of a speaker (i.e., their evaluative judgments) to a hearer. Such knowledge allows the hearer to better predict the speaker's behavior and, in cases where they disagree, make attempts to correct their beliefs. Therefore, having a strong formalism for learning from normative testimony is necessary to build socially competent AI systems. However, existing approaches have been mostly empirical, concerned with learning external responses (e.g., praise) [Hamid, N. H. A., et al., 2015; Sarathy, V. et al., 2017] rather than internal beliefs, and/or do not consider vital features like deontic ambiguity and speaker reliability (though Andrighetto, G. et al. [2010] do briefly discuss reliability).

This paper makes such a theoretical contribution and provides a philosophically justified formalism for learning a populations' normative beliefs from their normative testimony, considering both deontic ambiguity and speaker reliability. This work may be categorized as a novel approach to *norm synthesis* via group consensus [Kadir, M. R. A. and Selamat, A., 2018]. We start with a running example to build intuition around the process of learning from normative testimony. We then describe our representation scheme for normative beliefs and testimony. We then define a set of intuitive axioms of norm learning. Next, we construct our formal theory of norm learning, including philosophical justification and proof that it satisfies each axiom. Lastly, we theoretically evaluate its complexity and deontic consistency. We conclude with a discussion of related and future work.

2 Learning from Normative Testimony

Imagine you are visiting your friend Mary who just moved to a remote island, one which you have never visited. Let's call this society the Flarps. When you meet up with Mary at the Flarp library she leans over to you and whispers, "just so you know, you don't have to wear shoes in public." Nevertheless, you leave them on. Later that day at Mary's favorite coffee shop you witness the barista, Demarcus, get scolded by his boss. He then turns and yells at a customer, "put on your shoes!" And as you go about your day, you receive more normative testimony about wearing shoes in public.

Back at Mary's place and in for the night, you realize that you are quite certain that shoes are not required here, but you are unsure whether they are prohibited or just optional. Though Demarcus disagreed, he was under a bit of pressure, and many others said you could take them off. You have thus collected a set of normative testimony, fused this evidence in some fashion, and concluded the population's normative belief. This is the evidential process of descriptive ethics.

Many interesting questions arise about this process. Should the order in which you fuse the normative testimony matter? Should we weigh one more than another? We explore such questions here as we attempt to formalize norm learning. We start by detailing our formal representation scheme.

2.1 Formalizing Normative Concepts

Our language for representing deontic statuses is the Three-Fold Classification (TTC) of Deontic Logic [McNamara, P., 1996]. This consists of three mutually exclusive deontic statuses: *{Obligatory, Impermissible, and Optional}*, taken as "must be done", "must not be done", and "neither must nor must not be done" respectively. We also consider the weaker deontic concepts of *Permissible* as *{Obligatory, Optional}*, *Omissible* as *{Optional, Impermissible}*, and *NonOptional* as *{Obligatory, Impermissible}.*

We consider a norm as an internal normative belief. We modify existing formal representations for such mental attitudes [Olson, T. & Forbus, K., 2023; 2021; Santos, J. S. et al., 2017] to represent a set of agents, rather than individuals:

Definition 1 *(Normative Belief). A normative belief is a four-tuple $\langle A, B, C, D \rangle$ where:*

- *A is a set of agents e.g., {Mary, Demarcus};*
- *B represents the behavior (e.g., WearingShoes);*
- *C represents the context (e.g., InPublic);*
- *$D \in$ TTC is a deontic status (e.g., Omissible).*

This can be read as "in context C, the agent(s) A evaluate(s) behavior B as D." Next, we define normative testimony.

Definition 2 *(Normative Testimony). An instance of normative testimony is a five-tuple $\langle S, B, C, D, T \rangle$ where:*

- *S is the speaker (e.g., Mary);*
- *B represents the behavior (e.g., WearingShoes);*
- *C represents the context (e.g., InPublic);*
- *$D \in$ TTC is a deontic status (e.g., Omissible);*
- *T is the context of the testimony (e.g., AtLibrary).*

Given these structures, learning a population's normative beliefs can be formally viewed with a function that aggregates a set of normative testimony to yield a normative belief: $F(\{\langle \mathrm{S1}, \mathrm{B}, \mathrm{C}, \mathrm{D1}, \mathrm{T1} \rangle, \ldots \langle \mathrm{Sn}, \mathrm{B}, \mathrm{C}, \mathrm{Dn}, \mathrm{Tn} \rangle\}) \rightarrow \langle \mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D} \rangle$, where each agent $\mathrm{Sn} \in \mathrm{A}$.

Our main contribution here is formally characterizing this norm learning function F. We do so in a mathematical fashion and start with the axioms.

2.2 Norm Learning Axioms

We can decompose our aggregation function F into a recursion with a binary fusion operator as below. Where x and y are relevant sets of normative testimony,

$$F(x) = x, for\ a\ singleton\ x;$$
$$F(x \cup y) = F(x) \circledcirc F(y).$$

The binary function $\circledcirc$ essentially takes in two normative testimony and fuses them in some fashion to yield a combined normative belief. Considering our Flarps example, we present five intuitive axioms governing this function.

Axiom 1 *(Conjunctive Pooling). Fusing normative testimony considers the agreement in opinions: $x \circledcirc y \cong x \cap y$.*

When merging normative testimony to learn a population's beliefs we must consider the agreement between the claims of its individuals. The disagreement between Mary and Demarcus would make us less confident that the Flarps believe wearing shoes in public is omissible or obligatory individually. But if they agreed, then they would strengthen each other's claims. Thus, $\circledcirc$ should be viewed as a conjunctive pooling operation [Dubois, D., & Prade, H., 1992].

Axiom 2 *(Commutative). The result of fusing two normative testimony is the same, regardless of order: $x \circledcirc y = y \circledcirc x$.*

A temporal ordering of events is created while considering Mary and Demarcus's normative testimony. But we argue that, once the evidence is evaluated by the learner (interpreted, assessed for reliability, etc.), the order in which it is fused should not fundamentally matter. For example, all else being equal, encountering Demarcus before meeting up with Mary should yield the same normative belief of the Flarps.

Axiom 3 *(Associative). The result of repeatedly fusing normative testimony is the same regardless of how the body of evidence is partitioned: $(x \circledcirc y) \circledcirc z = x \circledcirc (y \circledcirc z)$.*

Norm learning is an online process and thus fusing a body of normative testimony should be associative (once the evidence is evaluated). This implies that fundamentally batch processes (e.g., LLMs like Delphi [Jiang, L. et al., 2021]) will not suffice for building true artificial social agents.

Axiom 4 *(Idempotent). Fusing an instance of normative testimony with itself has no effect: $x \circledcirc x = x$.*

Repeatedly considering the same instance of normative testimony should not have any effect on our certainty in the corresponding normative belief. For instance, ruminating on Mary's testimony should not make us more certain that the Flarps believe wearing shoes in public is omissible.

Axiom 5 *(Non-Dictatorial). Fusing two instances of normative testimony should preserve the opinion of both speakers (a formal definition is provided in later sections).*

When merging evidence from different sources, the hearer should not ignore any set of speakers. Because we wish to know what the Flarps as a whole believe, Mary's normative testimony should not completely override Demarcus's nor any other of the citizens. Thus, $\circledcirc$ should be non-dictatorial [Dalla Pozza, G. et al., 2011]. We note that this axiom does interact with conjunctive pooling, as the intersection ignores disagreement. We discuss how to handle this in later sections.

3 Normative Testimony as Evidence

With these axioms in place, we now construct our evidence fusion operator. We start by considering how to formally represent normative testimony as evidence.

The simplest option would be to view normative testimony as a frequentist would and consider the distribution of the explicit statements. If every single Flarp that day said, "you must wear shoes in public", then we would simply claim that the Flarps believe it is obligatory. Such a view seems to reduce normative testimony to the speaker's normative belief. However, this is clearly not the case as we can lie with our normative testimony. Demarcus's normative testimony seemed to arise purely due to his boss. A remorseless meat-eater might say, "eating animals is wrong" only because they are with vegans.

Our semantics for normative testimony instead seem to be that an instance of normative testimony probabilistically bears on the corresponding belief. Presumably, all of the features—the speaker's bodily expressions, context of the discourse—come into play when determining this measure of reliability. E.g., if Demarcus's boss was not around, then his normative testimony would hold much more weight. We thus need to represent such reliability measures in our function like so: where $P_n \in [0,1]$ is the reliability of the testimony, $F(\{P_1(\langle \mathrm{S1, B, C, D1, T1} \rangle), \ldots\ P_n(\langle \mathrm{Sn, B, C, Dn, Tn} \rangle)\}) \rightarrow \langle \mathrm{A, B, C, D} \rangle$.

It seems intuitive that these probabilities are at least determined by the freedom of expression of the discourse context. However, it is unclear how to ground such measures. How do we determine when others are telling the truth about their internal mental states? We cannot perceive minds like we can perceive dice rolls. We examine this question next.

3.1 Probability in Normative Testimony

We could ground our reliability measures in behavior and reduce normative testimony to the overt performance of demand actions. This would then be confirmable via perception in the same way that regular testimony is. For example, you could determine Mary's reliability by observing if she praises or condemns people who wear shoes. However, our normative beliefs are what we should really wish to teach our AI systems. And our beliefs and overt behavior are not always consistent [Borg J.S. et al., 2006; Buckholtz, J.W. et al., 2015; Gold, N. et al., 2015]. An agent may believe an action is impermissible and still not scold others for it (and vice versa).

So, if we stick with our intended cognitivist interpretation of normative testimony as assigning some degree of support for the claim "I *believe* this action is good/bad/neutral", then must the hearer verify this claim in some way? If so, how?

We argue that such probability measures for reliability can be grounded in meta-cognition and theory of mind. Under this interpretation, the higher the probability assigned to a *speaker's* reliability, the more the *hearer* correctly relays their own normative beliefs when in the speaker's position. Thus, each of us have probability measures for "how often I, the hearer, say my true normative beliefs in situations like this." We then assume "other people are like me" and project these probabilities as the reliability of the speaker (for a similar hypothesis, see Meltzoff [2007] and for a computational model of theory of mind, see Rabkina et al. [2017]). We formally define our hypothesis below.

Hypothesis 1 *(As Reliable as Me). Given a discourse context T, hearer H takes the chance that speaker S's normative testimony* $\langle S, B_1, C_1, D_1, T \rangle$ *relays their internal normative belief* $\langle \{S\}, B_1, C_1, D_1 \rangle$*, as the chance that their own normative testimony in context T* $\langle H, B_2, C_2, D_2, T \rangle$*, would relay their own normative belief* $\langle \{H\}, B_2, C_2, D_2 \rangle$*.*

By this hypothesis, our reliability measures for Mary and Demarcus' testimony are determined by how often we, the hearer, would be truthful about our normative beliefs in the Flarp library and when scolded by a boss respectively. With this philosophical grounding for such probabilities, we move on to continue formalizing our norm learning function.

4 Normative Testimony and Belief Functions

At this point our norm learning function F takes in probabilities on normative testimony and computes normative beliefs, likely with some measure of certainty. Such a framework of course aligns with a Bayesian approach. However, such a view requires data sources we do not have. How could we ever get prior and conditional probabilities for normative beliefs? Again, we cannot perceive people's mental states like we can observe dice rolls. We argue that epistemics do not fit into such a chance picture, and thus Bayesian semantics do not fit here either.

Instead, we argue for a relaxation of subjective probability theory, specifically belief functions [Shafer, G., 1976]. Belief functions emerge when we cannot fit our question into traditional probability theory, but we can justify probabilities for a related question. We then extend these subjective probabilities to get degrees of belief for the question we are truly interested in. And this framing fits quite nicely here. While we may only have probabilities for how reliable an agent's normative testimony may be, this is of course related to the question of what their normative beliefs actually are.

To illustrate, under this interpretation we would first consider the chance that Mary is providing truthful normative testimony in the Flarp library. Denote this as s. Again, s is determined by reflecting on the self and determining how truthful we, the hearer, would be in this context. This then serves as a premise along a chain of reasoning to the Flarp's normative belief. Thus, we get the syllogism below (inspired by Shafer's [1981] example due to J.H. Lambert).

- P1: A belief of an agent is a belief of their group.
- P2: Mary is a Flarp.
- *P3: Mary believes wearing shoes in public is omissible.*
- C: Flarps believe wearing shoes in public is omissible.

We, the hearer, assign probability s to premise three above. Therefore, our argument that "Flarps believe wearing shoes in public is omissible" is sound with chance s and unsound with chance $1 - s$. We argue that these are suitable semantics for normative testimony because while epistemics cannot fit into a chance picture, truth-telling can.

4.1 Background on Dempster-Shafer Theory

Here we specifically build upon Dempster-Shafer's theory of belief functions. AI researchers have applied this mathematical theory of evidence for sensor fusion [Premaratne, K. et al., 2009], learning indirect speech acts [Wen, R. et al., 2020], and more empirical norm learning work [Olson, T. and Forbus, K., 2021; 2023; Sarathy, V. et al., 2017]. We define the concepts of DS theory below.

Definition 3 *(Frame of Discernment). DS theory considers an exhaustive set called the frame of discernment (FOD), denoted as* Θ*, of elements that are mutually exclusive. We can interpret each element of* Θ *as an answer to our question.*

Definition 4 *(Mass Assignment). A mass assignment, or basic belief assignment (BBA), is a function, denoted as m, that maps each subset of Θ to a real number in $[0,1]$, such that $m(\emptyset) = 0$ and $\sum_{A \in 2^{\Theta}} m(A) = 1$. A mass assignment represents a judgment of the degree to which the evidence supports the set of propositions. Mass is assigned to non-singleton sets in cases of ambiguity or ignorance.*

Definition 5 *(Focal Element). The focal elements of a mass assignment are those sets with non-zero mass.*

Definition 6 *(Belief Function). Given set A, the belief function of $A = Bel(A) = \sum_{B|B \subseteq A} m(B)$.*

Definition 7 *(Plausibility Function). Given set A, the plausibility of $A = Pl(A) = \sum_{B|B\cap A \neq \emptyset} m(B)$. We can also compute this from belief as: $Pl(A) = 1 - Bel(A^C)$.*

5 Formalizing Norm Learning

Our approach to representing normative testimony as evidence in DS theory involves: 1) given the frame of discernment, convert normative testimony into an answer on this frame and 2) given a reliability measure, assign belief measures to this evidence, yielding a corresponding mass assignment. We define concepts relevant to this process below.

Definition 8 *(Deontic Frame of Discernment). A deontic frame of discernment $\Theta_{a,b,c}$ is the set of possible normative beliefs for a set of agents, a behavior, and a context: $\{\langle a, b, c, d\rangle | \; d \in TTC\}$.*

We use a few abbreviated notations for a frame. First, unless necessary, we omit certain features and write Θ. We denote specific subsets of a frame with a functional notation: $\Theta(D) = \{\langle \mathrm{a, b, c, d}\rangle | \; d \in D\}$ where $D \subseteq TTC$. We also abbreviate non-singleton subsets with corresponding weaker deontic status e.g., $\Theta(\{Opt, Imp\}) = \Theta(\{Omissible\})$.

Definition 9 *(Deontic Mass Assignment). A deontic mass assignment is a mass assignment on the subsets of a deontic frame. Formally, a deontic mass assignment is a mass assignment $m_{\{A\dots Z\}}: 2^{\Theta} \to [0,1]$ where Θ is a deontic frame of discernment and $A \dots Z$ are instances of normative testimony.*

Definition 10 *(Body of Evidence). Given deontic frame Θ, its body of evidence (BoE) is $BoE(\Theta) = \{m_X | \; m_X$ is a deontic mass assignment on $2^{\Theta}\}$.*

The ability to assign mass to subsets in DS theory, rather than just propositions, is powerful for norm learning as it naturally handles deontically ambiguous claims like Mary's. In this case, we assign mass to $\Theta_{A,B,C}(\{Omissible\})$, without needing to decide how to distribute the mass to each disjunct.

We have formalized normative testimony as evidence. We now discuss how to aggregate this evidence to learn from multiple sources i.e., we finally define our novel function F. Again, this involves chaining our binary fusion operator $\odot$ over the body of evidence and we have argued that this operator must satisfy five axioms. Thankfully, Dempster has provided us with a binary operator that comes quite close.

Definition 11 *(Dempster's Rule of Combination). Dempster's rule of combination (Dempster, 1967) computes the sum of the mass product intersections. It is defined as:*
$m_1 \oplus m_2(c) = \sum_{a\cap b=c} m_1(a)m_2(b)\,/(1-K) \quad \forall c \subseteq \Theta$
where m_1 and m_2 are two independent mass assignments on the same frame Θ, and conflict measure K is computed as: $K = \sum_{a\cap b=\emptyset} m_1(a)m_2(b)$. We use the following set notation for a fused assignment: $m_{X\cup Y}(c) \equiv m_X \oplus m_Y(c)$.

5.1 Example: The Norms of the Flarps

To illustrate Dempster's rule in our formalism, we revisit the Flarps. We wish to know how they evaluate wearing shoes in public. Thus, we are computing $F\big(BoE(\Theta)\big)$ where $\Theta =$ $\{\langle \mathrm{Flarps, WearingShoes, InPublic, Obligatory}\rangle,$
$\langle \mathrm{Flarps, WearingShoes, InPublic, Optional}\rangle,$
$\langle \mathrm{Flarps, WearingShoes, InPublic, Impermissible}\rangle\}$.

Because Mary is a Flarp, her normative testimony $e1$, assigns evidence on this frame, abbreviated as Θ. Assessed with a 0.9 degree of reliability, $e1$ produces the deontic mass assignment: $m_{\{e1\}}\big(\Theta(\{Omissible\})\big) = 0.9, m_{\{e1\}}(\Theta) = 0.1$.

We have another relevant piece of evidence, $e2$, when Demarcus yelled, "put on shoes!" Given the perceived pressure from his boss, we evaluated this as being much less reliable: $m_{\{e2\}}\big(\Theta(\{Obl\})\big) = 0.3, m_{\{e2\}}(\Theta) = 0.7$.

To determine the Flarps' normative belief, we then fuse these two pieces of evidence with Dempster's rule. This results in the mass product intersections shown in Table 1. Because Mary and Demarcus disagreed, there is mass assigned to the empty set. This is normalized out via $K = 0.27$, resulting in the fused mass assignment described below.
"Flarps believe wearing shoes in public is omissible"
$= m_{\{e1,e2\}}\big(\Theta(\{Om\})\big) = 0.63/(1-0.27) = 0.8631$;
"Flarps believe wearing shoes in public is obligatory"
$= m_{\{e1,e2\}}\big(\Theta(\{Obl\})\big) = 0.03/(1-0.27) = 0.0412$;
"Flarps believe it is of some deontic status" (ignorance)
$= m_{\{e1,e2\}}(\Theta) = 0.07/(1-0.27) = 0.0957$.

Because Mary's testimony was taken as much more reliable than Demarcus's, the model is still quite certain that the Flarps believe wearing shoes in public is omissible:
$Bel_{\{e1,e2\}}\big(\Theta(\{Om\})\big) = m_{\{e1,e2\}}\big(\Theta(\{Om\})\big) = 0.8631$;
$Pl_{\{e1,e2\}}\big(\Theta(\{Om\})\big) = m_{\{e1,e2\}}\big(\Theta(\{Om\})\big) + m_{\{e1,e2\}}(\Theta) =$ $0.8631 + 0.0957 = 0.9588$.

This example illustrates how norms are synthesized considering both deontic ambiguity and speaker reliability. Next we examine Dempster's rule with respect to our axioms.

	$m_{\{e1\}}\big(\Theta(\{Om\})\big)$ $= 0.9$	$m_{\{e1\}}(\Theta)$ $= 0.1$
$m_{\{e2\}}\big(\Theta(\{Obl\})\big)$ $= 0.3$	∅=0.27	*Θ({Obl})=0.03*
$m_{\{e2\}}(\Theta) = 0.7$	*Θ({Om})=0.63*	*Θ=0.07*

Table 1: Mary and Demarcus's Fused Mass Assignment (focal elements italicized and conflict underlined)

5.2 A Modified Fusion Rule

Dempster's rule satisfies most of our axioms. It is a conjunctive pooling operation, and it is both commutative and associative [Sentz, K. & Ferson, S., 2002]. However, as noted previously, being a conjunctive pooling operation, Dempster's rule is dictatorial in cases where one mass assignment is definitive, and another non-agreeing mass assignment is uncertain. It is also not idempotent by default. Here we provide a modified fusion rule that satisfies these two conditions. We first provide a formal definition for non-dictatorial.

Definition 12 *(Non-Dictatorial). We say that a fusion rule $\dot{R}$ is non-dictatorial when: $(\forall m_1, m_2, f)\, m_1 \dot{R} m_2(f) > 0$, where m_1 and m_2 are mass assignments on the same frame and $f \in$ focal elements of m_1.*

To illustrate that Dempster's rule is dictatorial, imagine Mary's testimony was taken to be definitive as $m_{\{e1\}}\big(\Theta(\{Om\})\big) = 1.0$. Because Demarcus's testimony does not agree (the intersection of $\{Obl\}$ and $\{Om\}$ is the empty set), it would be ignored. To avoid such a dictatorship, we must instead define our mass assignments correctly by assigning a non-zero mass to the frame. That is, a hearer must always leave a small portion of doubt in a speaker's claim. This is formalized with the following function, denoted with an overline. Where the amount of doubt $\varepsilon \in (0,1)$ (default of 0.1), m is a deontic mass assignment on frame Θ, $d \subseteq \Theta$,

$$\overline{m}(d) = \begin{cases} m(d),\ m(\Theta) > 0 \\ \varepsilon,\ d = \Theta \wedge m(\Theta) = 0 \\ m(d) \times (1 - \varepsilon),\ d \subset \Theta \wedge m(\Theta) = 0. \end{cases} \tag{1}$$

Before fusion, a deontic mass assignment must now first be processed by this function. Not only does this ensure there can be no dictator, but it also fixes counterintuitive cases that result from normalization like that provided by Zadeh [1984].

We then use set-theoretic operations to ensure fusion is idempotent. Given two deontic mass assignments, if one has already been fused with the other (conditions 1 and 2 in equation 2), then the mass assignment with the most accumulated evidence is returned. If the two are disjoint (condition 3), then they are fused with Dempster's rule. However, if they intersect but one does not subsume the other (condition 4), then fusion is undefined. With these modifications, our updated fusion rule ⊙ (or in short form with the set union "∪") is formalized below. Given two deontic mass assignments m_X, m_Y on deontic frame Θ,

$$m_X(c) \odot m_Y(c) \equiv m_{X \cup Y}(c) = \\ \begin{cases} \overline{m}_X(c),\ Y \subseteq X \\ \overline{m}_Y(c),\ X \subset Y \\ \overline{m}_X(c) \oplus \overline{m}_Y(c),\ X \cap Y = \emptyset \\ undefined,\ X \cap Y \neq \emptyset \wedge Y \nsubseteq X \wedge X \not\subset Y \end{cases} \\ \forall c \subseteq \Theta. \tag{2}$$

Our fusion rule ⊙ is non-dictatorial, in short, because each deontic mass assignment now has a non-zero mass assigned to the entire frame, and thus all focal elements have a non-zero mass product intersection when fused, as the intersection of a focal element and the entire frame is itself. It is idempotent because of conditions 1 and 2 of equation 2.

This fusion rule now satisfies all of our norm learning axioms. It is a conjunctive pooling operation (axiom 1) and is commutative (axiom 2), associative (axiom 3), idempotent (axiom 4), and non-dictatorial (axiom 5). We have thus axiomatically constructed our norm learning function F, defined below. Where x and y are subsets of a deontic frame's BoE,

$$F(x) = x, for\ a\ singleton\ x; \\ F(x \cup y) = F(x) \odot F(y). \tag{3}$$

We move on to discuss truth-conditions within our semantics and introduce a final axiom along with proof that our formalism satisfies it.

5.3 The Semantics of Normative Belief

While it is useful to have degrees of belief, when acting or predicting based on learned norms, a definitive claim needs to be made. For example, if we wish to not upset anyone, then to decide whether or not to wear shoes in Flarpville we must convert our degree of belief to a definitive norm of the Flarps. We thus need to define when a normative belief is true.

Intuitively, a normative belief is true when its body of evidence sufficiently supports it. We formally define this below as the mean of belief and plausibility (estimating the true support) being greater than or equal to 0.9. We arbitrarily chose this high threshold, but one can imagine defining a personality spectrum from gullible (→ 0.0) to skeptical (→ 1.0).

Given a normative belief $\mathcal{N} = \langle a, b, c, d \rangle$ with deontic frame Θ, and Bel and Pl functions stemming from the fused mass assignment $F(BoE(\Theta))$, we compute the truth of $\mathcal{N}$ as:

$$\begin{cases} True,\ \big(Bel\big(\Theta(\{d\})\big) + Pl\big(\Theta(\{d\})\big)\big)/2 \geq 0.9 \\ False,\ \big(Bel\big(\Theta(\{d\})\big) + Pl\big(\Theta(\{d\})\big)\big)/2 < 0.9. \end{cases} \tag{4}$$

For example, based on its body of evidence, the normative belief $\langle \mathrm{Flarps}, \mathrm{WearingShoes}, \mathrm{InPublic}, \mathrm{Omissible} \rangle$ is true, as the center of its belief and plausibility is greater than this threshold: $(0.8631 + 0.9588)/2 \geq 0.9$.

Under these semantics, inferences between deontic statuses should be sound. That is, our truth function above should not produce two deontically inconsistent normative beliefs (e.g., Mary can't believe wearing shoes in public is both omissible and obligatory). Disagreement should instead be represented with underlying uncertainty measures. This idea yields our sixth and final axiom of deontic consistency.

Axiom 6 *(Deontically Consistent). Given deontically inconsistent statuses D and E, the normative beliefs $\langle A, B, C, D \rangle$ and $\langle A, B, C, E \rangle$ should not both be true.*

Next, we analyze the complexity of computing normative beliefs and prove that it is deontically consistent.

6 Theoretical Evaluation

Querying for the truth of a normative belief involves gathering its BoE by A) looping through all evidence and for each,

B) comparing its agents, behavior, and context with those queried for. Then, C) fusing evidence with $F(BoE)$. Lastly, D) computing belief and plausibility values. The runtime of this algorithm is thus: $[\mathcal{O}(\mathrm{A}) \times \mathcal{O}(\mathrm{B})] + \mathcal{O}(\mathrm{C}) + \mathcal{O}(\mathrm{D})$. For analyzing this complexity, consider the parameters n as the magnitude of all evidence, m as the magnitude of the relevant BoE, d as the magnitude of our deontic frame.

Theoretically, n and m are unbounded (though practically limited by the lifetime of the organism). Thus, $\mathcal{O}(\mathrm{A}) = \mathcal{O}(\mathrm{n})$. The complexity of comparing a mass assignment's agents, behavior, and context with those queried for is bounded and thus $\mathcal{O}(\mathrm{B}) = \mathcal{O}(1)$. Therefore, the runtime of distilling the body of evidence is $[\mathcal{O}(\mathrm{A}) \times \mathcal{O}(\mathrm{B})] = \mathcal{O}(\mathrm{n}) \times \mathcal{O}(1) = \mathcal{O}(\mathrm{n})$. As the deontic frame is bounded by the size of TTC at $d = 3$, the complexity of fusing two deontic mass assignments is bounded by some constant c. Thus, our pair-wise fusion operation $F(BoE)$ (step C) involves $c \times (m-1)$ operations. Therefore, $\mathcal{O}(\mathrm{C}) = \mathcal{O}(m)$. Step D merely involves a summation over a finite set and thus, $\mathcal{O}(\mathrm{D}) = \mathcal{O}(1)$.

Considering these bounds, computing the truth value of a normative belief (equation 4) has a linear time complexity of: $[\mathcal{O}(\mathrm{A}) \times \mathcal{O}(1)] + \mathcal{O}(\mathrm{C}) + \mathcal{O}(1) = \mathcal{O}(\mathrm{n}) + \mathcal{O}(\mathrm{m})$, where n is the total amount of normative testimony and m is the amount of normative testimony relevant to our query.

Next, we show that by utilizing DS theory our semantics are deontically consistent. We also show how this can slightly reduce the runtime of computing truth values.

6.1 Deontic Consistency of Normative Belief

To save space, for a normative belief $\langle A, B, C, D \rangle$ we leave $A, B,$ and C implicit and abbreviate the belief as $B(D)$. We abbreviate the subset of its deontic frame as $\Theta(\{D\})$.

We first prove that if the threshold for $B(D)$ is defined above 0.5, then no two non-intersecting subsets of its deontic frame can both be believed. We note that this also holds in the other direction but omit the proof to save space.

Theorem 1 *(With a truth value threshold > 0.5, two inconsistent normative beliefs cannot both be true). Let $B(D)$ be true if and only if $Bel(\Theta(\{D\})) + Pl(\Theta(\{D\}))/2 > \alpha$. If $\alpha \geq 0.5$, then given $B(D)$ is true for deontic frame Θ, $\nexists \Theta(\{E\})$ s.t. $\Theta(\{E\}) \cap \Theta(\{D\}) = \emptyset$ and $B(E)$ is true.*

Proof. Let $\alpha \geq 0.5$. Let $B(X)$ be true if and only if $Bel(\Theta(\{X\})) + Pl(\Theta(\{X\}))/2 > \alpha$, or when $Bel(\Theta(\{X\})) + Pl(\Theta(\{X\})) > 2\alpha$. Let $\Theta(\{D\}), \Theta(\{E\}) \subseteq \Theta$ s.t. $\Theta(\{D\}) \cap \Theta(\{E\}) = \emptyset$. Assume $B(D)$ and $B(E)$ are true. Because $B(D)$ is true, rewriting plausibility in the ordinal relation, we get: $Bel(\Theta(\{D\})) + [1 - Bel(\Theta(\{D\})^C]) > 2\alpha \geq 1$. So, $Bel(\Theta(\{D\})) > Bel(\Theta(\{D\})^C)$. Similarly, $Bel(\Theta(\{E\})) > Bel(\Theta(\{E\})^C)$. Because $\Theta(\{D\}) \cap \Theta(\{E\}) = \emptyset$, $\Theta(\{D\}) \subseteq \Theta(\{E\})^C$ and $\Theta(\{E\}) \subseteq \Theta(\{D\})^C$. By definition of the belief function and transitivity of subsets, $Bel(\Theta(\{D\})) \leq Bel(\Theta(\{E\})^C)$. Similarly, $Bel(\Theta(\{E\})) \leq Bel(\Theta(\{D\})^C)$. Lemma 1.1 follows.

Lemma 1.1 *For any $X \subseteq Y \subseteq \Theta$, $Bel(X) \leq Bel(Y)$.*

We now have the following ordinal relations: $Bel(\Theta(\{D\})) > Bel(\Theta(\{D\})^C) \geq Bel(\Theta(\{E\}))$ and $Bel(\Theta(\{E\})) > Bel(\Theta(\{E\})^C) \geq Bel(\Theta(\{D\}))$. Thus, both $Bel(\Theta(\{D\})) > Bel(\Theta(\{E\}))$ and $Bel(\Theta(\{E\})) > Bel(\Theta(\{D\}))$, which is a contradiction. Thus, $B(D)$ and $B(E)$ cannot both be true. □

To illustrate, theorem 1 holds that with a threshold of truth above 0.5, two normative beliefs like "Flarps believe wearing shoes in public is obligatory" and "Flarps believe wearing shoes in public is impermissible" cannot both be true.

Utilizing theorem 1, we now prove that certain relations of deontic logic are sound. For each theorem below, let the truth of a normative belief $B(D)$ be defined as previously (equation 4) with a threshold of 0.9.

Theorem 2. *The following syntactic entailments between normative beliefs in contrary deontic statuses are sound:*

$$B(OBL) \vdash \neg B(IMP)$$
$$B(OBL) \vdash \neg B(OPT)$$
$$B(IMP) \vdash \neg B(OPT).$$

Proof. Let $B(OBL)$ be true. Because $\Theta(\{OPT\}) \cap \Theta(\{OBL\}) = \emptyset$, $\Theta(\{IMP\}) \cap \Theta(\{OBL\}) = \emptyset$, and our threshold $0.9 > 0.5$, by theorem 1, $\neg B(OPT)$ and $\neg B(IMP)$ are true. Without loss of generality, the same inference holds between all core deontic statuses. □

Theorem 3. *The following syntactic entailments between normative beliefs in contradictory deontic statuses are sound:*

$$B(OBL) \vdash \neg B(OM)$$
$$B(OPT) \vdash \neg B(NONOPT)$$
$$B(IMP) \vdash \neg B(PERM).$$

Proof. Let $B(OBL)$ be true. By definition, $\Theta(\{OM\}) = \Theta(\{OPT, IMP\})$ and thus $\Theta(\{OM\}) \cap \Theta(\{OBL\}) = \emptyset$. It follows from theorem 1 that $\neg B(OM)$ is true. Without loss of generality, the same relation holds from IMP to $PERM$, and OPT to $NONOPT$. □

Theorem 4. *The following syntactic entailments between subsumed normative beliefs are sound:*

$$B(OBL) \vdash B(PERM)$$
$$B(OBL) \vdash B(NONOPT)$$
$$B(IMP) \vdash B(OM)$$
$$B(IMP) \vdash B(NONOPT)$$
$$B(OPT) \vdash B(PERM)$$
$$B(OPT) \vdash B(OM).$$

To aid in proving theorem 4, we first prove theorem 5.

Theorem 5 *(If normative belief with deontic status D is true, then given E subsumes D, normative belief with deontic status E is true). Where $\Theta(\{D\}) \subseteq \Theta(\{E\})$, if $B(D)$, then $B(E)$.*

Proof. Let $B(D)$ be true, where $\Theta(\{D\}) \subseteq \Theta(\{E\})$. By lemma 1.1, $Bel(\Theta(\{D\})) \leq Bel(\Theta(\{E\}))$. Because $\Theta(\{D\}) \subseteq \Theta(\{E\})$, all sets that intersect with $\Theta(\{D\})$ also intersect with $\Theta(\{E\})$. Lemma 5.1 follows.

Lemma 5.1. *For any $X \subseteq Y \subseteq \Theta$, $Pl(X) \leq Pl(Y)$.*

By lemma 5.1, $Pl(\Theta(\{E\})) \geq Pl(\Theta(\{D\}))$. Because $B(D)$ is true, $Bel(\Theta(\{D\})) + Pl(\Theta(\{D\}))/2 \geq 0.9$. Thus, $Bel(\Theta(\{E\})) + Pl(\Theta(\{E\}))/2 \geq 0.9$ as well. Therefore, $B(E)$ is true. □

We now utilize theorem 5 to prove theorem 4.

Proof. Let $B(OBL)$ be true. As defined, $\Theta(\{OBL\}) \subseteq \Theta(\{PERM\}), \Theta(\{NONOPT\})$. Thus, by theorem 5,

$B(PERM)$ and $B(NONOPT)$ are also true. Without loss of generality, the same inference holds from belief in all deontic statuses to deontic statuses that subsume them. □

We have shown that certain syntactic inferences of standard deontic logic are sound given our semantics. This means that no two inconsistent normative beliefs can be learned and thus our formal theory is deontically consistent (satisfies axiom 6). This is practically useful as it ensures consistency when using learned beliefs to predict behavior. This also means that unless there is new evidence, we can simply prove normative beliefs rather than computing their truth value, reducing complexity. For instance, if it is true that "X is viewed as obligatory", then we can prove "X is viewed as permissible" via theorem 4, rather than calculating its truth value.

Note that we omit inheritance principles OB-RM (as described by Ross [1941]) or, in conditional form, CPI [Olson, T. and Forbus, K., 2023], from this calculus as they are unsound. To illustrate, consider a population that believes harming an agent is impermissible and that war entails harm. Thus, via inheritance it could be proven that the population also believes that war is impermissible. However, one can clearly imagine a model in which this is not true. We humans are not always consistent across such entailments and thus such inheritance principles *should* be unsound in epistemic formalisms. However, because they aim at ideal practical reasoning, such rules could be used to point out and then correct the non-rational normative beliefs of a population.

7 Related Work

The work presented here differs from existing norm learning/synthesis approaches that are grounded in *external* factors. That a sufficient number of individuals demand compliance with a norm, that it comes from an authority, has negative consequences, and so on (as in [Sarathy, V. et al., 2017] and [Hamid, N. H. A., et al., 2015]), are empirical conditions that can easily be verified via scientific method. However, verifying *internal* mental states is not as simple (As Reliable as Me hypothesis). This idea has been of particular interest in law, with respect to expert testimony on mental states and insanity defenses [Levy, K. M., 2019; Slobogin, C., 2008]. Recent neural approaches [Freitas dos Santos, T. et al., 2023; Jiang et al., 2021] differ for similar reasons. Also, though they can perform quite well empirically on tasks, being black boxes, their formal properties are difficult to examine.

In previous work [Olson and Forbus, 2021] we presented a norm learning implementation that also utilized DS theory. However, we left a lot of theoretical work undone. By working from axioms and formal definitions here, we have made fundamental improvements including a new fusion rule and a sound deontic calculus. But the biggest difference again lies in our probability semantics. Previously, probabilities represented "moral trust". However, such a model could not learn the norms of an immoral population, as their trust measures would be too low. If the goal is to learn prescriptive ethics (moral truths), then this is quite valuable. But here we are concerned with descriptive ethics. We wish to learn norms whether or not they are morally correct. We note that this should not be used prescriptively to guide the actions of AI systems without explicit moral guardrails (e.g., Olson and Forbus [2023]). Instead, this can be used descriptively to predict behavior and correct immoral beliefs.

8 Conclusion and Future Work

While it is recognized that the ability to learn the normative beliefs of others is necessary for social competence, there exists few formal models of this process. Those approaches that do exist learn only external demand responses (e.g., praise), have been mostly empirical, and/or have ignored the uncertainty of internal mental states. This paper thus makes a theoretical contribution by providing a philosophically justified formalism for learning internal normative beliefs from normative testimony. We have provided a set of axioms for this process. We have provided an interpretation of probability measures as being grounded in the hearer's mental model. We have built an improved binary fusion operator and shown that it satisfies the axioms. Lastly, we have provided a complexity analysis of the norm learning algorithm and a sound deontic calculus for reasoning about learned norms.

We admit that hypothesis 1 introduces a sincere "cold start problem", as artificial agents lacking such knowledge cannot estimate reliability measures. This may be solvable by starting agents with needs and desires similar to ours, but this remains an open research question.

Though our work here is theoretical, it is also of practical use. As we have discussed, this could be implemented to learn norms of a population to then predict their behavior or change their beliefs. This will be an important capability for future artificial social agents. This could also be implemented for learning and comparing beliefs of subpopulations (e.g., left vs right wing ideals) considering reliability. In such applications, the reliability of each opinion depends on the context in which it is provided (hence the appeal of secret ballots).

In future work we plan to examine the potential for prejudice in our model, as it currently clumps agents who have not provided evidence in with others that have. To remove this feature, the algorithm would ensure each agent in the query has provided normative testimony. We also plan to explore strengthening the independence requirement and only consider each agent's most recent normative testimony. This would make the model fairer as a single agent could no longer flood it with evidence.

Acknowledgments

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. We are grateful to Kyla Ebels-Duggan for discussions that helped shape ideas in this work. We also thank the anonymous reviewers for their feedback and suggestions.

References

[Andrighetto, G. et al., 2010] Giulia Andrighetto, Marco Campennì, Federico Cecconi, and Rosaria Conte. The complex loop of norm emergence: A simulation model. In Simulating Interacting Agents and Social Phenomena:

The second world congress: 19-35. Springer Japan. doi.org/10.1007/978-4-431-99781-8_2. 2010.

[Awad, E. et al., 2018] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. “The moral machine experiment.” *Nature* 563, no. 7729: 59-64. 2018.

[Borg J.S. et al., 2006] Jana Schaich Borg, Catherine Hynes, John Van Horn, Scott Grafton, and Walter Sinnott-Armstrong. “Consequences, action, and intention as factors in moral judgments: An fMRI investigation.” *Journal of cognitive neuroscience* 18, no. 5: 803-817. 2006.

[Buckholtz, J.W. et al., 2015] Joshua W. Buckholtz, Justin W. Martin, Michael T. Treadway, Katherine Jan, David H. Zald, Owen Jones, and Rene Marois. “From blame to punishment: disrupting prefrontal cortex activity reveals norm enforcement mechanisms.” *Neuron* 87, no. 6: 1369-1380. 2015.

[Dalla Pozza, G. et al., 2011] Giorgio Dalla Pozza, Maria Silvia Pini, Francesca Rossi, and K. Brent Venable. “Multi-agent soft constraint aggregation via sequential voting.” In *Twenty-Second International Joint Conference on Artificial Intelligence*. 2011.

[Dempster, A.P. 1967] Arthur P. Dempster. Upper and Lower Probabilities Induced by a Multivalued Mapping. In Annals of Mathematical Statistics 38, no. 2, 325-339. doi:10.1214/aoms/1177698950. 1967.

[Dubois, D., & Prade, H., 1992]. On the combination of evidence in various mathematical frameworks. In Reliability data collection and analysis (pp. 213-241). Dordrecht: Springer Netherlands. 1992.

[Freitas dos Santos, T. et al., 2023] Thiago Freitas dos Santos, Nardine Osman, and Marco Schorlemmer. “A multi-scenario approach to continuously learn and understand norm violations.” *Autonomous Agents and Multi-Agent Systems* 37, no. 2: 38. 2023.

[Gold, N. et al., 2015] Natalie Gold, Briony D. Pulford, and Andrew M. Colman. “Do as I say, don’t do as I do: Differences in moral judgments do not translate into differences in decisions in real-life trolley problems.” *Journal of Economic Psychology* 47: 50-61. 2015.

[Hamid, N. H. A., et al., 2015] Nurzeatul Hamimah Abdul Hamid, Mohd Sharifuddin Ahmad, Azhana Ahmad, Aida Mustapha, Moamin A. Mahmoud, and Mohd Zaliman Mohd Yusoff. "Trusting norms: A conceptual norms’ trust framework for norms adoption in open normative multi-agent systems." In Distributed Computing and Artificial Intelligence, 12th International Conference, pp. 149-157. Springer International Publishing, 2015.

[Hills, A., 2009] Alison Hills. Moral Testimony and Moral Epistemology. Ethics, 120(1), 94–127. doi.org/10.1086/648610. 2009.

[Jiang, L. et al., 2021] Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. “Delphi: Towards machine ethics and norms.” *arXiv preprint arXiv:2110.07574* 6. 2021.

[Kadir, M. R. A. and Selamat, A., 2018] Mohd Rashdan Abdul Kadir and Ali Selamat. A Categorization of Runtime Norm Synthesis in Normative Multi-Agent Systems. In 2018 IEEE Conference on e-Learning, e-Management and e-Services (IC3e): 128-133. IEEE. 2018.

[Levy, K. M., 2019] Ken M. Levy. Normative Ignorance: A Critical Connection Between the Insanity and Mistake of Law Defenses. Fla. St. UL Rev., 47, 411. 2019.

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

[Meltzoff, A.N., 2007] Andrew N. Meltzoff. 'Like me': a foundation for social cognition. Dev Sci. Jan;10(1):126-34. doi: 10.1111/j.1467-7687.2007.00574.x. PMID: 17181710; PMCID: PMC1852489. 2007.

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

[Olson, T. and Forbus, K., 2023] Taylor Olson and Kenneth Forbus. Mitigating Adversarial Norm Training With Moral Axioms. In Proceedings of AAAI-23. 2023.

[Olson, T., 2024] Towards Unifying the Descriptive and Prescriptive for Machine Ethics. In P. Wu, M. Salpukas, H.F. Wu, S. Ellsworth (Eds.), Trolley Crash: Approaching Key Metrics for Ethical AI Practitioners, Researchers, and Policy Makers. (Chapter 5). Elsevier. 2024.

[Premaratne, K. et al., 2009] Kamal Premaratne, Manohar N. Murthi, Jinsong Zhang, Matthias Scheutz, and Peter H. Bauer. A Dempster-Shafer theoretic conditional approach to evidence updating for fusion of hard and soft data. In Proceedings of the 12th International Conference on Information Fusion (pp. 2122-2129). 2009.

[Rabkina, I. et al., 2017] Irina Rabkina, Clifton McFate, Kenneth D. Forbus, and Christian Hoyos. Towards a Computational Analogical Theory of Mind. In Proceedings of the 39th Annual Conference of the Cognitive Science Society, 2949-2954. 2017

[Ross, A., 1941] Alf Ross. Imperatives and Logic (in Discussions). Theoria, 7(1): 53–71. doi:10.1111/j.1755-2567.1941.tb00034.x. 1941.

[Santos, J. S. et al., 2017] Santos, Jéssica S., Jean O. Zahn, Eduardo A. Silvestre, Viviane T. Silva, and Wamberto W. Vasconcelos. “Detection and resolution of normative conflicts in multi-agent systems: a literature survey.” *Autonomous agents and multi-agent systems* 31: 1236-1282. 2017.

[Sarathy, V. et al., 2017] Vasanth Sarathy, Matthias Scheutz, Yoed N. Kenett, Mowafak Allaham, Joseph L. Austerweil, and Bertram F. Malle. Mental representations and computational modeling of context-specific human norm systems. In Proceedings of the 39th Annual Meeting of the Cognitive Science Society. 2017.

[Sentz, K. & Ferson, S., 2002] Karl Sentz and Scott Ferson. “Combination of Evidence in Dempster-Shafer Theory”. United States. doi.org/10.2172/800792. 2002.

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

[Shafer, G., 1981] Glenn Shafer. Constructive probability. Synthese, 1-60. 1981.

[Slobogin, C., 2008] Christopher Slobogin. Experts, mental states, and acts. Seton Hall L. Rev., 38, 1009. 2008.

[Wen, R. et al., 2020] Ruchen Wen, Mohammed Aun Siddiqui, and Tom Williams. Dempster-Shafer Theoretic Learning of Indirect Speech Act Comprehension Norms. Proceedings of the AAAI Conference on Artificial Intelligence (pp. 10410-10417). 2020.

[Zadeh, L.A., 1984] Lotfi A. Zadeh. Review of a mathematical theory of evidence. AI magazine, 5(3), 81-81. 1984.