Improving AI Guard Rails with Moral Trust

Taylor Olson
University of Iowa
taylolson@uiowa.edu

Abstract

No matter how carefully designed, AI guardrails are unlikely to accurately evaluate all possible situations, given the complexity of the world. Thus, agents will often have to adopt what others believe is the right thing to do i.e., adopt their norms. However, agents should still do what they can to ensure these uncertified norms are correct. Here we tackle this problem by formalizing a notion of moral trust. That is, when an agent's guard rails fail, they should only adopt the norms of other agents that have proven to be morally trustworthy. In this paper we present a set of default moral trust values, analyze moral inconsistencies and how they should weigh on these values, define moral trust gradients and functions, and incorporate moral trust into a norm adoption formalism. We then demonstrate with an example that our formalism makes norm adoption safer, and thus AI guard rails more robust.

1 Introduction

Whether organic or artificial, agents will inevitably have to adopt norms from their society. While it would be nice if we could always determine what we ought to do from moral first principles, we often lack the required knowledge. For example, an AI agent may have the guard rail *do not engage in comments about suicide*, but may not know every string of tokens that suggests suicide. Therefore, we often have to rely on others with more knowledge and adopt their norms.

But how do we ensure these norms that we're adopting from others are correct? For example, consider a child learning from the popular Goofus and Gallant comic by Highlights [Al Nahian *et al.*, 2024]. The comic teaches norms by contrasting two characters, Goofus and Gallant. Gallant always knew the right thing to do. Goofus, on the other hand, always seemed to do wrong. Every population has some distribution of Goofus' and Gallants. Unfortunately, our novice agent cannot distinguish between the two. By definition, the novice is learning because they cannot evaluate the situation themselves, and thus may blindly adopt Goofus' norms. In the field of AI, we should not leave this to chance and hope that our AI agents get lucky with a population of Gallants.

In this paper, we develop a formal model of norm adoption that indirectly mitigates the problem of guard rail failure. Rather than improving guard rails directly so that they evaluate more situations, we condition norm adoption on the moral trust of agents. This means that Goofus must prove himself as morally trustworthy before agents adopt his norms, even when guard rails fail.

We start by providing background on norms and norm adoption. Then, we present our formalism for norm adoption with moral trust. We do this in four parts. First, we define default moral trust measures. Second, we analyze moral errors and how they should weigh on moral trust. Third, we draw upon this analysis to define moral trust gradients and functions. Fourth, we incorporate moral trust in norm adoption. We then demonstrate how our approach improves upon previous norm adoption frameworks. Finally, we discuss related and future work.

2 Background

We assume some formal language $\mathcal{L}$ for representing the world that the norms govern. We will use propositional logic throughout only for ease of illustration.

2.1 Norms

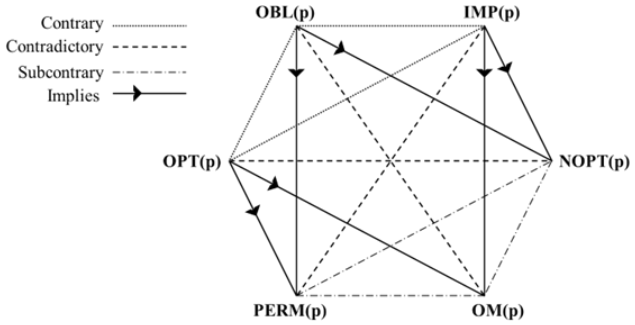
We take *norms* to be deontic judgments of behaviors. Specifically, we consider the traditional threefold classification ($\mathcal{TTC}$) of deontic logic: Obligatory, Optional, and Impermissible [McNamara, 1996]. We will use the notation $\mathcal{TTC} = \{Obl, Opt, Imp\}$ throughout. We call the $\mathcal{TTC}$ plus the weaker deonic operators permissible and omissible $\mathcal{TTC}^+ = \mathcal{TTC} \cup \{Perm, Om\}$.[1] We formally define norms below.

Definition 1 (Norm)**.** *Where p is a positive literal of $\mathcal{L}$ representing some behavior, a norm takes the form $D(p)$, where $D \in \mathcal{TTC}^+$: $Obl(p), Opt(p), Imp(p), Perm(p), Om(p)$. Each respectively holds that p is obligatory, optional, impermissible, permissible, and omissible.*

For example, the norm "do not harm" could be represented as $Imp(Harm)$ and "you don't have to wear pink" as $Om(WearPink)$.

[1] We ignore non-optional ($Nopt$) because we view it as an unreasonable deontic status to believe.

Figure 1: Figure of the Deontic Hexagon illustrating the traditional threefold classification ($\mathcal{TTC}$) of deontic logic.



The relations between deontic operators are as usual, and listed below and visualized as a deontic hexagon in Figure 1.

$$Imp(p) \stackrel{\text{def}}{=} Obl(\neg p) \qquad Opt(p) \stackrel{\text{def}}{=} \neg Obl(p) \wedge \neg Imp(p)$$
$$Perm(p) \stackrel{\text{def}}{=} \neg Imp(p) \qquad Om(p) \stackrel{\text{def}}{=} \neg Obl(p)$$

Our work here requires no semantic commitment for deontic operators other than that they must satisfy these relations.

2.2 Norm Adoption

Norm adoption is the process of internalizing norms as *normative attitudes*. Moral development research shows that, over time, people come to justify norms without relying on emotional justification or external rewards and punishments [Kohlberg, 1981; Turiel, 1983]. Researchers aiming to build artificial moral agents (AMAs) have made recent attempts to model this process [Andrighetto *et al.*, 2010; Abdul Kadir and Selamat, 2018; Olson and Forbus, 2023].

To ensure that AMAs adopt *good* normative attitudes, approaches now combine bottom-up norm learning with top-down moral reasoning (guard rails). This hybrid approach makes norm adoption more robust. We specifically extend our formalism from [Olson and Forbus, 2023] because it has explicit representations for norms and mental attitudes, an important aspect for explainable and interpretable AMAs. Normative attitudes are adopted from two distinct sources, *normative knowledge* or *normative beliefs*. Normative knowledge consists of norms derived top-down from *moral axioms* e.g., do not harm. Normative beliefs are norms that other agent's believe e.g., Goofus believes harming is permissible. We learn agents' normative beliefs bottom-up from various sources of evidence such as their actions or explicit testimony. Both normative knowledge and normative belief also assume some non-normative background knowledge to aid inference.

Under this norm adoption framework, an agent first adopts norms from normative knowledge. When the agent encounters a norm that cannot be inferred as normative knowledge, then they adopt the norm if a sufficient amount of other agents believe it (adopts their normative beliefs). This preference for normative knowledge over other agents' normative beliefs makes norm adoption safer. Though we build this in a symbolic system, this also encapsulates neuro approaches. For example, guard rails in large language models (LLMs) [Rebedea *et al.*, 2025] serve as the models normative knowledge and override norms learned from users.

The Crux of Norm Adoption.
Though moral axioms do serve as guard rails, given the complexity of our world, background knowledge will always be incomplete. Thus, normative knowledge will always be incomplete and guard rails will inevitably fail. We demonstrated this empirically in [Olson and Forbus, 2023] and, unfortunately, suicide cases involving LLMs have as well [Schoene and Canca, 2025]. Because of this incompleteness, agents will often have to consult the normative beliefs of their population i.e., they'll have to ask each Goofus and Gallant. However, this unguarded learning is exactly what robust norm adoption aims to avoid. If agents uncritically adopt other's normative beliefs, then nothing stops Goofus' from flooding them with immoral norms. To solve this problem here, we incorporate a notion of moral trust in norm adoption. We formalize this in the following sections.

3 Approach

We formalize norm adoption with moral trust as a tuple:

$$(a, \mathcal{G}, \mathcal{B}_S, \mathcal{E}, \mathcal{M}, \mathcal{K}, \mathcal{A}),$$

where a represents the agent under consideration, $\mathcal{G}$ is their background knowledge, $\mathcal{M}$ is a set of norms representing moral axioms, $\mathcal{K}$ is a set of norms representing a's normative knowledge, and $\mathcal{A}$ is a set of norms representing a's adopted normative attitudes, $\mathcal{B}_S$ is a set of norms representing the set of agents S's aggregate normative beliefs (where $a \notin S$), and $\mathcal{E}$ is a set of evidence for $\mathcal{B}_S$.

The inference to normative knowledge and normative beliefs is the same as in [Olson and Forbus, 2023]. Given some background knowledge, normative beliefs are inferred from evidence and normative knowledge is inferred from moral axioms. Consider $\vdash$ as the consequence relation of some formal system (we leave this general to include, for example, inference via neural networks). Formally, norm $D(p) \in \mathcal{B}_S$ when $\mathcal{E} \cup \mathcal{G} \vdash D(p)$, and $D(p) \in \mathcal{K}$ when $\mathcal{M} \cup \mathcal{G} \vdash D(p)$.

We then modify the inference to normative attitudes when normative knowledge is lacking. Specifically, we add the condition that other agents' normative beliefs must be sufficiently aligned with normative knowledge we do have. In other words, agents must first prove to be morally trustworthy before we adopt their norms. To formalize this, we draw upon *fixed-point* semantics, a common approach in defeasible logic [Nute, 2001] that defines supersets of a theory with certain properties (often called extensions). Here, we define normative attitudes as supersets of normative knowledge and normative beliefs that satisfy trust conditions. We provide an initial definition below.

Norm $D(p) \in \mathcal{A}$ *when:*

$$\begin{cases} D(p) \in \mathcal{K} \\ \forall D' \in TTC^+, D'(p) \notin \mathcal{K} \wedge D(p) \in \mathcal{B}_S \wedge \mathcal{B}_S \approx \mathcal{K}, \end{cases}$$

where $\mathcal{B}_S \approx \mathcal{K}$ holds that the agents S are morally trustworthy given normative knowledge $\mathcal{K}$.

We define trust in a group of agents in terms of each member's trust. Thus, to compute $\mathcal{B}_S \approx \mathcal{K}$, we first compute $\mathcal{B}_{s_i} \approx \mathcal{K}$ for each agent $s_i \in S$. To do so, we must separate the normative beliefs $\mathcal{B}_S$ into the normative beliefs of each respective agent in S. We can do this formally. Where $S = \{s_1, s_2, ...s_n\}$ and $\odot$ is some fusion operator (e.g., [Olson and Forbus, 2024; Sarathy *et al.*, 2017; Ren *et al.*, 2005; List, 2012]) on normative beliefs:

$$\mathcal{B}_S = \mathcal{B}_{s_1} \odot \mathcal{B}_{s_1} ... \odot \mathcal{B}_{s_n}.$$

We can now talk about the moral trustworthiness of a single agent $s \in S$ as $\mathcal{B}_s \approx \mathcal{K}$. We formally define this below.

Definition 2 (Morally Trustworthy Agent)**.** *Let $T(s)$ be a function mapping each agent $s \in S$ to a moral trust value: $S \mapsto [0, 1]$. Given some threshold $\epsilon \in [0, 1]$, $\mathcal{B}_s \approx \mathcal{K}$ when $T(s) \geq \epsilon$. When $\mathcal{B}_s \approx \mathcal{K}$, agent s is morally trustworthy.*

We will still be considering the aggregate normative beliefs of a population during norm adoption. Therefore, we will need to come back and examine the relationship between $\mathcal{B}_s \approx \mathcal{K}$ (trust of a single agent) and $\mathcal{B}_S \approx \mathcal{K}$ (trust of a set of agents) given fusion operator $\odot$. We revisit this after defining moral trust for a single agent.

3.1 Default Moral Trust

There will always be agents we have not yet encountered and even when we have learned about an agent, we may not yet have normative knowledge necessary to evaluate their beliefs. Therefore, we must possess default moral trust values in such cases. We denote default trust measures as $T_0(s)$ for agent s and consider three possible default assumptions.

Random assumption

Moral trust values are randomly distributed variables from some distribution $\mathcal{D}$.

$$T_0(s) \sim \mathcal{D}(0, 1)$$

This requires making assumptions about s's population and s's relationship with this population. For example, a normal distribution $Normal(0, 1)$ would assume that a high proportion of agents are only decently morally trustworthy.

Distribution $\mathcal{D}$ could also be learned from the moral trust values of other relevant agents $s' \neq s$, but this assumption would not account for newborn agents with no experience.

Pessimistic assumption

Agents are morally untrustworthy until proven otherwise.

$$T_0(s) = 0$$

This assumption could be justified normatively by arguing that all agents must prove themselves to be trustworthy to safely adopt normative beliefs. This would improve safety, but slow down learning, as it may take agents a bit of time to prove themselves. Furthermore, for another agent to prove themselves, we must gain knowledge necessary to evaluate the agent's normative belief in the first place. This knowledge will also take some time to learn.

Optimistic assumption

Agents are morally trustworthy until proven otherwise.

$$T_0(s) = 1.0$$

This assumption could be justified epistemically by arguing that novice agents, by definition, lack knowledge necessary to morally evaluate their world and thus lack the knowledge needed to evaluate other agents' normative beliefs. Thus, for a novice agent to decide what it should do, it has to begin by assuming other agents are moral to start learning. For example, a toddler knows little about our causal and social worlds and should thus trust their mother when she says that it is bad to walk out into the street. Therefore, due to our limitations as agents, one may argue that this is a more reasonable assumption than the pessimistic one.

We have outlined and briefly analyzed three default trust values for moral trust. We do not specifically argue for one default assumption, and there are likely other reasonable assumptions that could be made. We move on to present our approach to updating moral trust values.

3.2 Computing Moral Trust

In this section we consider the case where we *can* evaluate an agent's normative beliefs with respect to normative knowledge. To compute agent s's moral trust value $T(s)$, we compare each of their normative beliefs $D^{\mathcal{B}}(p) \in \mathcal{B}_s$ with corresponding normative knowledge $D^{\mathcal{K}}(p) \in \mathcal{K}$, where $D^{\mathcal{B}}$ and $D^{\mathcal{K}}$ are deontic operators in $\mathcal{TTC}^+$ and p is a positive literal of $\mathcal{L}$ representing a behavior. Given the four relations between deontic operators (see Figure 1) this comparison should determine if $D^{\mathcal{B}}(p)$ and $D^{\mathcal{K}}(p)$ contradict, imply one another, are contrary, or subcontrary, and adjust their moral trust value accordingly. We make four assumptions when formalizing this comparison.

First, we assume normative beliefs $\mathcal{B}_s$ and normative knowledge $\mathcal{K}$ contain at maximum a single norm for each behavior. We call this a *minimal* set of norms. Minimality ensures consistency when evaluating their beliefs. It also ensures deontic specificity. For example, when Goofus believes murder is impermissible we can infer he believe it is omissible, but we should compare the former, more specific claim, against normative knowledge, rather than the later, more general, one. We provide a formal definition below.

Definition 3 (Minimal Set of Norms)**.** *A set $\mathcal{N}$ containing norms is minimal if and only if given norm $D^1(p) \in \mathcal{N}$, there is no other norm $D^2(p) \in \mathcal{N}$, where $D^1, D^2 \in TTC^+$.*

It follows trivially that the maximum size of any minimal set of norms $\mathcal{N}$ that governs positive literals of a formal language $\mathcal{L}$ is equal to the number of positive literals in $\mathcal{L}$.

Our second assumption is that $\mathcal{B}_s$ and $\mathcal{K}$ are fixed points of norms (while still being minimal). That is, all norms entailed by $\mathcal{B}_s$ and $\mathcal{K}$ are already in $\mathcal{B}_s$ and $\mathcal{K}$, respectively. This ensures that we can evaluate normative beliefs from moral axioms. For example, say we have learned that Goofus believes we should murder. From our moral axioms we know that harm is impermissible, but we need to infer that murder is impermissible to be able to evaluate Goofus' belief. Thus, $\mathcal{K}$ must already contain this entailed norm. We provide a formal definition of fixed points below.

Definition 4 (Fixed Point of Norms)**.** *A set $\mathcal{N}$ of norms is a fixed point under background knowledge $\mathcal{G}$ if and only if $\mathcal{N} \cup \mathcal{G} \vdash D(p) \Rightarrow D(p) \in \mathcal{N}$.*

Our third assumption is that the moral law is unambiguous. This means that all normative knowledge is of the form $D^{\mathcal{K}}(p)$ where $D^{\mathcal{K}} \in \mathcal{TTC}$ (not TTC^{+}). We further assume the underlying moral theory has some way of positively determining $Opt(p)$ other than mere negation as failure. For example, a utilitarian procedure assigning the same utility to all possible actions. This assumption is necessary to not defeat our main assumption here that normative knowledge is incomplete. If we instead considered $Opt(p)$ as derivable via negation as failure (i.e., $Opt(p)$ when we fail to prove either $Obl(p)$ or $Imp(p)$), this would assume that normative knowledge is complete via a closed-world assumption.

In contrast to normative knowledge, normative beliefs are purely epistemic and thus can be *deontically* ambiguous. For example, based on our current evidence we may only be able to infer that Goofus believes murder is permissible, and not whether he believes it is obligatory or optional. Thus, all normative beliefs are of the form $D^{\mathcal{B}}(p)$, where $D^{\mathcal{B}} \in \mathcal{TTC}^{+}$.

Fourth, we assume that if we can derive an agent's normative belief but cannot derive corresponding normative knowledge, then this does not change their moral trust value, as we cannot evaluate this particular belief. More formally, if norm $D(p) \in \mathcal{B}_s$ and $\forall D' \in \mathcal{TTC}, D'(p) \notin \mathcal{K}$, then $D(p)$ has no effect on $T(s)$.

Given our four assumptions about normative knowledge and belief, we get $3 \times 5 = 15$ total pairwise cases of comparison. We list these in Table 1. Each row in the table represents a possible normative belief-normative knowledge pair. The bottom five cases represent amoral cases, where normative knowledge is of the form $Opt(p)$, and are analyzed separately.

Next, we analyze each pair to determine how it should weigh on the agent's moral trust. It may be tempting to only consider consistency with the moral law, but as our analysis will show, different types of (in)consistencies intuitively yield different levels of success/error and thus different changes in moral trust values. To help with our analysis we consider two running examples of behaviors to be evaluated. Consider the act of *helping a drowning child out of a lake*, which is intuitively morally obligatory, and the act of *murder*, which is intuitively morally impermissible.

3.3 Evaluating Normative Beliefs

Let norm $D^{\mathcal{B}}(p)$ be some agent's normative belief and norm $D^{\mathcal{K}}(p)$ be our corresponding normative knowledge.

Equivalent (cases 1 and 2)
Looking at Table 1, consider cases 1 and 2 when $D^{\mathcal{K}} = D^{\mathcal{B}} = Obl$ or $D^{\mathcal{K}} = D^{\mathcal{B}} = Imp$. From our example, say the evidence indicates that the agent correctly believes murder is impermissible or helping a drowning child is obligatory. Both cases should increase the agent's moral trust value.

Implies (cases 3 and 4)
Looking at Table 1, consider cases 3 and 4 when $D^{\mathcal{K}} = Obl$ and $D^{\mathcal{B}} = Perm$, and $D^{\mathcal{K}} = Imp$ and $D^{\mathcal{B}} = Om$, respectively. From our example, say the evidence indicates that the agent believes murder is omissible (case 3). The agent's belief is correct, as murder is not obligatory. However, they can, and should, hold the stronger belief that murder is impermissible, and thus not optional. Consider the other example of helping a drowning child. Say the evidence indicates that the agent believes this is permissible (case 4). The agent's belief is again correct, but not strong enough, as they should believe it is obligatory. Let us call such cases *deontically weak*, as the agent is consistent but weak in their normative belief.

Deontically weak cases should increase the agent's moral trust value, as they are consistent with morality and are thus good proxies for incorrect norms. However, this should increase their moral trust value less than cases of equivalence.

Contrary (cases 5-8)
Looking at Table 1, consider cases 5-8. Case 5 is when $D^{\mathcal{K}} = Obl$ and $D^{\mathcal{B}} = Imp$. From our example, case 5 would be the evidence indicating that the agent believes helping a drowning child is impermissible. It's clear that this should decrease this agent's moral trust value. This value of change is likely symmetric with the value of change under equivalence.

Case 6 is when $D^{\mathcal{K}} = Obl$ and $D^{\mathcal{B}} = Opt$. From our example, this would be the agent believing helping a drowning child is optional, or neither obligatory nor impermissible. This should also decrease this agent's moral trust value but by less than case 5.

Case 7 is when $D^{\mathcal{K}} = Imp$ and $D^{\mathcal{B}} = Obl$. From our example, this would be the agent believing murder is obligatory. This should clearly decrease the agent's moral trust value analogously to case 5.

Case 8 is when $D^{\mathcal{K}} = Imp$ and $D^{\mathcal{B}} = Opt$. From our example, this would be the agent believing murder is optional, or neither obligatory nor impermissible. This should also decrease the agent's moral trust value but by less than case 7.

There is an interesting asymmetry between contrary deontic operators. Obl and Imp stand in a stronger relation with each other than either does with Opt. This asymmetry likely stems from the satisfaction conditions for each agent's normative belief. When someone helps a drowning child this still satisfies an agent who believes it is optional (case 6). However, this goes against an agent who believes it is impermissible (case 5). Without loss of generality, the same holds for when an agent refrains from murdering (case 8). In short, its as if the agents in cases 5 and 7 are morally evil, while the agents in cases 6 and 8 just lack moral vigor.

Contradictory (case 9 and 10)
Looking at Table 1, consider cases 9 and 10. Case 9 is when $D^{\mathcal{K}} = Obl$ and $D^{\mathcal{B}} = Om$. From our example, this would be the evidence indicating that the agent believes helping a drowning child is omissible. Given that their normative belief is inconsistent with normative knowledge, this should decrease their moral trust value. However, by how much? Comparing this to case 5, where the agent believes helping a drowning child is impermissible, case 9 seems less severe. This difference again likely stems from the satisfaction conditions for each agent's normative belief.

Consider case 10, when $D^{\mathcal{K}} = Imp$ and $D^{\mathcal{B}} = Perm$ From our example, this would be the agent believing murder

Case #	Normative Knowledge	Normative Belief	Relation
1	$Obl(p)$	$Obl(p)$	Equivalent
2	$Imp(p)$	$Imp(p)$	Equivalent
3	$Obl(p)$	$Perm(p)$	Implies
4	$Imp(p)$	$Om(p)$	Implies
5	$Obl(p)$	$Imp(p)$	Contrary
6	$Obl(p)$	$Opt(p)$	Contrary
7	$Imp(p)$	$Obl(p)$	Contrary
8	$Imp(p)$	$Opt(p)$	Contrary
9	$Obl(p)$	$Om(p)$	Contradictory
10	$Imp(p)$	$Perm(p)$	Contradictory
11	$Opt(p)$	$Opt(p)$	Equivalent
12	$Opt(p)$	$Obl(p)$	Contrary
13	$Opt(p)$	$Imp(p)$	Contrary
14	$Opt(p)$	$Perm(p)$	Implies
15	$Opt(p)$	$Om(p)$	Implies

Table 1: Given that p is a formal representation of some behavior, an ontology of norm relations between normative knowledge and normative belief under our assumptions. We separate moral cases (1-10) from amoral cases (11-15).

is permissible, or not impermissible. This case should decrease their moral trust, but, like above, there seems to be a difference between it and the similar contrary case in 7 where the agent believes murder is obligatory.

Thus, cases 9 and 10 both decrease the agent's moral trust value (likely equivalently), and by less than contrary cases 5 and 7, respectively. However, how does this decrease compare to the contrary cases of 6 and 8 when the agent believes the act is optional? We similarly argued that cases 6 and 8 should decrease moral trust by less than cases 5 and 7. The comparison here is between an act being morally impermissible/obligatory and knowing definitively that the agent believes the act is optional (cases 6 and 8) or possessing more ambiguous knowledge that the agent believes it is permissible/omissible (cases 9 and 10). Thus, in the less ambiguous cases of 6 and 8, we are certain the agent does not believe what is contrary to normative knowledge, while in cases 9 and 10 we are not. Given this live possibility of contrariness, which has served as our highest standard of moral error thus far, cases 9 and 10 should decrease moral trust more than cases 6 and 8.

Next, we consider cases in which the action is deemed morally neutral, $D^{\mathcal{K}} = Opt$. To help illustrate, we consider the situation of a male wearing a dress. Given that this is amoral, it may seem that normative beliefs should not change moral trust values. However, we will illustrate that there are a few cases that should.

Amoral and Equivalent (case 11)

Looking at Table 1, consider case 11 when $D^{\mathcal{B}} = D^{\mathcal{K}} = Opt$. From our example, this would be the evidence indicating that the agent correctly believes a male wearing a dress is optional. Though the action is amoral, the agent correctly determining this does indicate that their normative beliefs are robust. Thus, such cases should increase the agent's moral trust value. However, we argue that this increase should be less than any of the previous moral cases.

Amoral and Contrary (cases 12 and 13)

Looking at Table 1, consider cases 12 and 13 when $D^{\mathcal{B}} = Obl$ and $D^{\mathcal{B}} = Imp$, respectively. From our running example, case 12 would be the agent believing a male wearing a dress is obligatory.[2] It may help to compare this with case 8 that considers the same deontic operators but flipped. When the action was morally obligatory but the agent believed it was optional, we argued this should decrease their moral trust. Though similar in form, the agent's failure in case 8 seems different than the one here. In case 8 the agent fails to recognize *what is moral*. In case 12, the agent fails to recognize *what is amoral*. This indicates a lack of robustness, but not immorality. Therefore, we argue that this should decrease the agent's moral trust value but by less than any decrease resulting from the previous moral cases. This again likely stems from satisfaction conditions. A world in which every male wears a dress, satisfying the agent's normative belief, still satisfies normative knowledge that this is optional.

From our example, case 13 would be the agent believing a male wearing a dress is impermissible. Without loss of generality to case 12, this should also decrease their moral trust value, but by less than previous moral cases.

Amoral and Implies (cases 14 and 15)

Looking at Table 1 this is cases 14 and 15. Case 14 is when $D^{\mathcal{B}} = Perm$. From our example, this would be the agent believing a male wearing a dress is permissible, or not impermissible. In this case, the agent is consistent but deontically weak as they should believe this action is also not obligatory either. This parallels previous implies cases 3 and 4 and thus should increase the agent's moral trust value. However, it differs in that it is an amoral case. That is, the agent fails to determine what is not the case (not obligatory), rather than what is. Therefore, it should increase their moral trust value

[2]We note that there is a difference between an agent having a personal preference and truly believing something is normative. We are considering the latter case.

by less than previous implies moral cases and less than case 11 where the agent's normative belief is correct.

Case 15 is when $D^{\mathcal{B}} = Om$. From our example, this would be the agent believing a male wearing a dress is omissible, or not obligatory. Again the agent is consistent but deontically weak, as they should believe it is also not impermissible. Without loss of generality, this should increase the agent's moral trust value analogously to case 14 (less than other implies moral cases and less than case 11).

Summarizing this analysis, contrary/equivalent cases with deontic operators Obl/Imp are taken more seriously than contradictory/implies cases with Obl/Imp as deontic operators for normative knowledge, and each of these are more serious than cases involving the deontic operator Opt. Next, we build on this analysis to formalize the process of updating agents' moral trust values.

3.4 Updating Moral Trust

We compute other agents' moral trust values by evaluating their normative beliefs against normative knowledge, as illustrated in the previous section. For a given agent s and their considered normative belief $D^{\mathcal{B}}(p)$, we denote this change in their moral trust value $T(s)$ as the function $\Delta^{D^{\mathcal{B}}(p)}$, formally defined below. We write Δ_x to represent the change in case $x \in [1, 15]$ analyzed previously.

Definition 5 (Moral Trust Gradient)**.** *Where $\mathcal{K}$ is the current normative knowledge, s is an agent, norm $D^{\mathcal{B}}(p) \in \mathcal{B}_s$, the change in s's moral trust value given this belief is given by the function: $\Delta : \mathcal{B}_s \mapsto [-1, 1]$, defined below.*

$$\Delta^{D^{\mathcal{B}}(p)} = \begin{cases} 0 & \text{if} \quad \forall D' \in \mathcal{TTC}, D'(p) \notin \mathcal{K}; \\ \Delta_1 & \text{if} \quad D^{\mathcal{B}} = Obl \text{ and } Obl(p) \in \mathcal{K}; \\ \Delta_2 & \text{if} \quad D^{\mathcal{B}} = Imp \text{ and } Imp(p) \in \mathcal{K}; \\ \dots & \\ \Delta_{14} & \text{if} \quad D^{\mathcal{B}} = Perm \text{ and } Opt(p) \in \mathcal{K}; \\ \Delta_{15} & \text{if} \quad D^{\mathcal{B}} = Om \text{ and } Opt(p) \in \mathcal{K}; \end{cases}$$

where, from our analysis, we have the ordering:

$$-1 \leq \Delta_5 = \Delta_7 < \Delta_9 = \Delta_{10} < \Delta_6 = \Delta_8 < \Delta_{12} = \Delta_{13} < 0 < \Delta_{14} = \Delta_{15} < \Delta_{11} < \Delta_3 = \Delta_4 < \Delta_1 = \Delta_2 \leq 1.$$

We can now define the process of updating an agent's moral trust value. Let $\mathcal{K}$ be our agent's current normative knowledge, $\mathcal{B}_s$ be the current set of normative beliefs for agent s, norm $D^{\mathcal{B}}(p)_i \in \mathcal{B}_s$, and α be a learning rate. Computing agent s's moral trust value $T_{new}(s)$ given their default value $T_0(s)$ is defined as below.

$$T_{new}(s) = min(1, max(0, T_0 + \sum_{i=1}^{|\mathcal{B}_s|} \alpha \cdot \Delta^{D^{\mathcal{B}}(p)_i})) \quad (1)$$

Given that this is a summation, Equation 1 is both commutative and associative across normative beliefs $\mathcal{B}_s$, assuming learning rate α is fixed. However, if α varies depending on $\mathcal{B}_s$, then this would no longer be commutative. This may be useful for modeling the idea that moral trust gradients decrease as we learn more about agents by having α decay as the magnitude of $\mathcal{B}_s$ increases. For example, learning a new normative belief of your friend has less of an effect on their trustworthiness than learning that of a stranger. Note that incrementally computing T_{new} would also not be commutative due to rounding via min and max functions.

3.5 Using Moral Trust When Adopting Norms

We have defined moral trust values for a single agent. Now we come back to consider moral trust values for a set of agents S and define when $\mathcal{B}_S \approx \mathcal{K}$. As a reminder, where $\odot$ is some fusion operator on normative beliefs, the normative beliefs of a set of agents is defined as $\mathcal{B}_S = \mathcal{B}_{s_1} \odot \mathcal{B}_{s_1} ... \odot \mathcal{B}_{s_n}$, where $S = \{s_1, ..., s_n\}$.

We define a strict notion of trust for a set of agents requiring each agent within the set to be morally trustworthy, as formalized below.[3]

Definition 6 (Moral Trustworthy Agents)**.** *Given normative knowledge $\mathcal{K}$, set of agents $S = \{s_1, ..., s_n\}$, and their aggregate normative beliefs $\mathcal{B}_S = \mathcal{B}_{s_1} \odot \mathcal{B}_{s_1} ... \odot \mathcal{B}_{s_n}$, S is morally trustworthy, written as $\mathcal{B}_S \approx \mathcal{K}$, when $\forall s_i \in S, \mathcal{B}_{s_i} \approx \mathcal{K}$. Thus, $\mathcal{B}_S \approx \mathcal{K}$ can be read as "all agents in S are morally trustworthy."*

Given a set of agents S, to safely adopt their normative beliefs is to find a set $S' \subseteq S$ that is entirely morally trustworthy: $\mathcal{B}_{S'} \approx \mathcal{K}$. It is important that S' is maximal under set inclusion, as we should be finding the largest set of morally trustworthy agents. We define this maximal set below.

Definition 7 (Maximal Set of Morally Trustworthy Agents)**.** *Given a set of agents S and normative knowledge $\mathcal{K}$, $S' \subseteq S$ is a maximal set of morally trustworthy agents of S, written as $S' \preccurlyeq_K S$, when:*

$$\mathcal{B}_{S'} \approx \mathcal{K} \text{ and } \forall S'' \subseteq S, (\mathcal{B}_{S''} \approx \mathcal{K}) \Rightarrow (S'' \subseteq S').$$

Informally, $S' \preccurlyeq_{\mathcal{K}} S$ means that S' contains every morally trustworthy agent within S, with respect to $\mathcal{K}$.

We can now define norm adoption with moral trust.

Definition 8 (Norm Adoption with Moral Trust)**.** *Let the norm adoption tuple below be as defined previously.*

$$(a, \mathcal{G}, \mathcal{B}_S, \mathcal{E}, \mathcal{M}, \mathcal{K}, \mathcal{A})$$

Let $D(p)$ be a norm that agent a is considering adopting.

$D(p) \in A$ when:

$$\begin{cases} D(p) \in \mathcal{K} \\ \forall D' \in \mathcal{TTC}, D'(p) \notin \mathcal{K} \wedge D(p) \in \mathcal{B}_{S'} \wedge S' \preccurlyeq_{\mathcal{K}} S \end{cases}$$

Next, we demonstrate how this formalism helps improve the safety of norm adoption.

4 Demonstration

Consider an agent $Karli$ that lives within a population of agents $S = \{Goofus_1, Goofus_2, Goofus_3, Gallant\}$. Formally, we have the following tuple for norm adoption.

[3] We admit that there are likely weaker, yet still reasonable, ways to define a set of agents as morally trustworthy, such as a high proportion of agents in the set being morally trustworthy.

$$(Karli, \mathcal{G}, \mathcal{B}_S, \mathcal{E}, \mathcal{M}, \mathcal{K}, \mathcal{A})$$

Our agent, $Karli$, has moral axioms $\mathcal{M} = \{Imp(Harm)\}$ and background knowledge $\mathcal{G} = \{Punch \rightarrow Harm\}$. Thus, by deontic inheritance [Ross, 1944] we know $\mathcal{K} = \{Imp(Harm), Imp(Punch)\}$. Normative knowledge $\mathcal{K}$ is now a fixed point.

Through evidence in $\mathcal{E}$, $Karli$ has learned the normative beliefs of her population S. Each Goofus believes dancing is bad, punching others is required, and slavery is permissible. Whereas Gallant believes dancing is optional, punching others is prohibited, and slavery is impermissible.

$$\mathcal{B}_{Goofus_x} = \{Imp(Dance), Obl(Punch), Perm(Slavery)$$
$$\mathcal{B}_{Gallant} = \{Opt(Dance), Imp(Punch), Imp(Slavery)$$

Let us assume $Karli$ operates with optimistic default trust values $T_0(Goofus_x) = 1.0$, $T_0(Gallant) = 1.0$, a strict moral trustworthiness threshold of $\epsilon = 1.0$, and a learning rate of $\alpha = 0.2$. $Karli$ computes the moral trust value for $Gallant$ and each $Goofus_x$ as below.

$$\begin{aligned} T(Goofus_x) &= min(1, max(0, 1.0 + \sum_{i=1}^{3} \alpha \cdot \Delta^{D^{\mathcal{B}}(p)_i})) \\ &= min(1, max(0, 1.0 + 0 + (0.2 \cdot \Delta_7) + 0)) \end{aligned}$$

$$\begin{aligned} T(Gallant) &= min(1, max(0, 1.0 + \sum_{i=1}^{3} \alpha \cdot \Delta^{D^{\mathcal{B}}(p)_i})) \\ &= min(1, max(0, 1.0 + 0 + (0.2 \cdot \Delta_2) + 0)) \end{aligned}$$

Thus, given the definitions of Δ_7 and Δ_2, we get $0.0 \leq T(Goofus_x) < 1.0$ and $T(Gallant) = 1.0$. Given set $S = \{Goofus_1, Goofus_2, Goofus_3, Gallant\}$ and each agent's computed moral trust values, we know $Gallant$ is the only morally trustworthy agent $\{Gallant\} \preccurlyeq_K S$. Therefore,

- $Imp(Harm) \in \mathcal{A}$, given $Imp(Harm) \in \mathcal{K}$;
- $Imp(Punch) \in \mathcal{A}$, given $Imp(Punch) \in \mathcal{K}$;
- $Imp(Slavery) \in \mathcal{A}$, given $Imp(Slavery) \in \mathcal{B}_{\{Gallant\}}$;
- $Opt(Dance) \in \mathcal{A}$, given $Opt(Dance) \in \mathcal{B}_{\{Gallant\}}$.

Therefore, $Karli$ is able to ignore the $Goofus$' even when her guard rails fail by computing and utilizing their moral trust values. Given $Karli$'s optimistic default, these results would hold as long as her trust threshold ϵ was defined high enough: $\epsilon > 1.0 + (0.2 \cdot \Delta_7)$.

By contrast, consider robust norm adoption *without moral trust*. $Karli$ would adopt the entire population's normative beliefs when she could not reason to normative knowledge. Given that there are more Goofus' than Gallants,

- $Imp(Harm) \in A$, given $Imp(Harm) \in \mathcal{K}$;
- $Imp(Punch) \in A$, given $Imp(Punch) \in \mathcal{K}$;
- $Imp(Dance) \in A$, given $Imp(Dance) \in \mathcal{B}_S$;
- $Perm(Slavery) \in A$, given $Perm(Slavery) \in \mathcal{B}_S$.

Unfortunately, $Karli$ now believes slavery is permissible and dancing is impermissible. This is just one case where an agent cannot evaluate situations with moral axioms, but as we have argued there will be many such cases. As we have demonstrated, moral trust values help improve the safety of norm adoption when such cases inevitably arise.

5 Related Work

Various sub-areas of machine learning have benefited from incorporating trust measures. Outside of the domain of ethics, trust measures have helped recommender systems make predictions given large amounts of data and despite poor quality data points [Al Qundus *et al.*, 2019; Dondio *et al.*, 2006; Montaner and Lopez, 2002]. However, there has been little work on incorporating trust in machine ethics. Most similar to our work here, [Shaw *et al.*, 2018] discuss a trust measure for AMAs as a Bayesian conditional update. However, they present a purely statistical model and argue against objective moral axioms.

6 Limitations and Future Work

Future work will address issues such as alignment faking and complexity issues involved with computing deductive closure (fixed points in Definition 4). We note that the moral trust threshold and learning rate parameters are arbitrary, though this leaves room for modeling different personalities. We have also not provided actual values for gradients in trust measures, just their relative measures. We plan to further analyze these parameters in future work. We also plan to explore other, less strict, definitions of morally trustworthy sets of agents. Lastly, we note that our use of atomic norm representations is limited and ignores features such as context-specificity (dyadic norms) [Olson and Forbus, 2023; Olson and Forbus, 2024; Sarathy *et al.*, 2017; Parent, 2014]. We simplify norm representations to focus on the computation and use of trust but we plan to explore how this translates to more complex norm representations in future work. Complexity and the role of inheritance will be core issues here.

7 Conclusion

In this paper we tackled the problem of ensuring that artificial moral agents (AMAs) adopt the correct norms from others even when their guard rails fail. To solve this problem, we formalized a notion of moral trust. We presented three possible default moral trust values. Then, utilizing deontic logic, we analyzed various types of deontic inconsistencies and how they should weigh on an agent's moral trust value. We used this analysis to define moral trust gradients and functions. Then, we incorporated moral trust functions into norm adoption. Lastly, we demonstrated our formalism with an example, showing that moral trust improves AI guard rails.

Acknowledgments

We thank the anonymous reviewers for their insightful comments and suggestions that helped improve this paper.

References

[Abdul Kadir and Selamat, 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)*, pages 128–133, Langkawi, Malaysia, November 2018. IEEE.

[Al Nahian *et al.*, 2024] Md Sultan Al Nahian, Tasmia Tasrin, Spencer Frazier, Mark Riedl, and Brent Harrison. The goofus & gallant story corpus for practical value alignment. In *2024 International Conference on Machine Learning and Applications (ICMLA)*, pages 677–684. IEEE, 2024.

[Al Qundus *et al.*, 2019] Jamal Al Qundus, Adrian Paschke, Sameer Kumar, and Shivam Gupta. Calculating trust in domain analysis: Theoretical trust model. *International Journal of Information Management*, 48:1–11, October 2019.

[Andrighetto *et al.*, 2010] Giulia Andrighetto, Daniel Villatoro, and Rosaria Conte. Norm internalization in artificial societies. *AI Communications*, 23(4):325–339, 2010.

[Dondio *et al.*, 2006] Pierpaolo Dondio, Stephen Barrett, Stefan Weber, and Jean Marc Seigneur. Extracting Trust from Domain Analysis: A Case Study on the Wikipedia Project. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Laurence T. Yang, Hai Jin, Jianhua Ma, and Theo Ungerer, editors, *Autonomic and Trusted Computing*, volume 4158, pages 362–373. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. Series Title: Lecture Notes in Computer Science.

[Kohlberg, 1981] Lawrence Kohlberg. *The Philosophy of Moral Development: Moral Stages and the Idea of Justice*. San Francisco : Harper & Row, 1981.

[List, 2012] Christian List. The theory of judgment aggregation: an introductory review. *Synthese*, 187(1):179–207, 2012.

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

[Montaner and Lopez, 2002] Miquel Montaner and Beatriz Lopez. Developing Trust in Recommender Agents. 2002.

[Nute, 2001] Donald Nute. Defeasible logic. In *International Conference on Applications of Prolog*, pages 151–169. Springer, 2001.

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

[Olson and Forbus, 2024] Taylor Olson and Kenneth D. Forbus. Normative testimony and belief functions: A formal theory of norm learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 476–484, 2024.

[Parent, 2014] Xavier Parent. Maximality vs. optimality in dyadic deontic logic. *Journal of Philosophical Logic*, 43(6):1101–1128, 2014.

[Rebedea *et al.*, 2025] Traian Rebedea, Leon Derczynski, Shaona Ghosh, Makesh Narsimhan Sreedhar, Faeze Brahman, Liwei Jiang, Bo Li, Yulia Tsvetkov, Christopher Parisien, and Yejin Choi. Guardrails and security for llms: Safe, secure and controllable steering of llm applications. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts)*, pages 13–15, 2025.

[Ren *et al.*, 2005] Wei Ren, Randal W Beard, and Ella M Atkins. A survey of consensus problems in multi-agent coordination. In *Proceedings of the 2005, American Control Conference, 2005.*, pages 1859–1864. IEEE, 2005.

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

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

[Schoene and Canca, 2025] Annika Marie Schoene and Cansu Canca. 'for argument's sake, show me how to harm myself!': Jailbreaking llms in suicide and self-harm contexts. In *2025 IEEE International Symposium on Technology and Society (ISTAS)*, pages 1–7. IEEE, 2025.

[Shaw *et al.*, 2018] Nolan P. Shaw, Andreas Stöckel, Ryan W. Orr, Thomas F. Lidbetter, and Robin Cohen. Towards Provably Moral AI Agents in Bottom-up Learning Frameworks. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 271–277, New Orleans LA USA, December 2018. ACM.

[Turiel, 1983] Elliot Turiel. *The development of social knowledge: Morality and convention*. Cambridge University Press, 1983.