Towards a Model of the Dynamics of Norm-guided Blaming

Emily LeBlanc
U.S. Naval Research Laboratory
emily.leblanc@nrl.navy.mil

Abstract

In this paper, we propose a novel framework for modeling how people assign blame for norm-violating conduct, taking into consideration the effects of both the violator's mental state and history of prior violations on the amount of blame they receive. We rely on the action language $\mathcal{AL}$ for the representation of domains and their evolution over time.

1 Introduction

All collaborative behavior among humans involves expectations, or *norms*, that people have of one another. These expectations arise from a commonly accepted set of rules and preferences (e.g., duties, roles, social conventions) that serve as guidelines for efficient and ethical conduct. Norm-guided behavior plays an extremely important role in the success of human societies, and is therefore a necessary condition of the future artificial agents who will be integrated into them.

An important aspect of norm-guided behavior involves people monitoring one another for norm-violating conduct and engaging in a blame-oriented interactions with one another when violations are detected. Recent research into the psychology of blaming tells us that social expectations for "fair" judgments of blame (i.e., supported by evidence) motivate people to carefully consider any available information about a norm violator's causal contribution and mental states [Malle *et al.*, 2014]. Moreover, specific types of qualitative information (e.g., the violator's intentions, reasons for acting, or capacity to prevent) have distinct effects on the amount of blame[1] that people assign for violations [Monroe and Malle, 2019]. Consider the following example adapted from the cited work in which someone has violated a standard social norm concerning respectful treatment of other's property:

Scenario 1. *Neil fried the motherboard on Allen's computer. He forgot to ground his tools when he was working on the computer and shorted a circuit.*

It can be reasoned from the first sentence that Neil violated the norm by damaging Allen's computer. With this information alone, one may assign an initial amount of blame to Neil based on the severity of the violation. The second sentence, however, provides additional information that enables us to infer that Neil acted unintentionally, but that the violation was preventable. As a result, the amount of blame assigned to Neil is likely to decrease slightly relative to the initial judgment, but less so than if Neil was incapable of preventing the damage (e.g. the tools were grounded, but the computer was wired incorrectly).

An individual's history of norm-violating behavior has also been shown to influence how people assign blame, with the amount increasing for repeated violations of the same or similar kind [Robinson and Darley, 2019]. For instance, there is typically an expectation that someone who has already received blame for an unintentional violation will recognize their obligations and/or maximize their capacity to prevent the same or similar outcomes in the future [Monroe and Malle, 2019]. Returning to Example 1, if Neil fries a second motherboard due to once again neglecting to ground the tools, he is likely to receive increased blame because he should have "learned" from his initial violation and adjusted his behavior.

We propose a novel theoretical framework for modeling norm-guided blaming that captures the effects of both blame-relevant mental state information and past violations on the amount of blame that an agent receives for a norm violation. The framework is built upon the action language $\mathcal{AL}$ [Baral and Gelfond, 2000], which enables us to characterize critical aspects of norm-guided blaming for complex, dynamic domains. $\mathcal{AL}$ can be conveniently translated to the declarative programming language ASP [Balduccini and Gelfond, 2003].

The remainder of the paper is organized as follows. In the next section, we review the psychological theories of norm-guided blaming underlying our approach. Following that, we provide background for our formalization of knowledge and events. Next, we present an overview of our proposed framework. We conclude with a brief discussion of directions for future work.

2 Psychology of Norm-guided Blaming

The Path Model of Blame [Malle *et al.*, 2014] describes a temporal ordering by which people predictably process specific types of blame-relevant information to arrive at a carefully reasoned judgment. According to this model, observing a norm violation prompts people to first determine whether a person is responsible for the outcome. If so, they next use

[1] In [Monroe and Malle, 2019], the "amount of blame" ranges from *no blame at all* to *the most blame one would ever give*.

New Information Condition	SRBP
Intentional only	↑↑
Intentional with bad reasons	↑↑↑
Intentional with good reasons	↓↓↓
Unintentional only	↓↓
Unintentional but preventable	↓
Unintentional and unpreventable	↓↓↓

Table 1: Blame change patterns with respect to initial judgments, as predicted by the Socially Regulated Blame Perspective (SRBP).

available information to assess whether the person acted intentionally or unintentionally. If they find that the action was intentional, then the amount of assigned blame depends on the person's (good or bad) reasons for acting. If their action was unintentional, then the amount of assigned blame depends on whether or not they were able (or obligated) to prevent the violation. Figure 1 depicts the blame-relevant concepts and processing paths of the Path Model that lead to a judgment of blame. Each node corresponds to the blamer's evaluation of whether or not that criteria has been met, and each edge represents a transition to either the next concept in the model or arrival at a judgement.

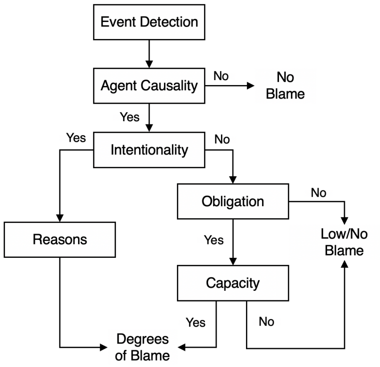

Figure 1: The Path Model of Blame: Blame-relevant concepts and processing paths (Malle, Guglielmo, & Monroe, 2014).

The Socially Regulated Blame Perspective [Monroe and Malle, 2019] arose from the Path Model of Blame as a set of predictions accounting for the effects of specific types of blame-relevant information on the amount of blame that people assign to someone who has violated a norm. Table 1 presents a reproduction of identified blame change patterns in response to distinct types of blame-relevant information. In the table, magnitudes of change in blame intensity are indicated by arrows (↑ for increase; ↓ for decrease) and the number of arrows indicates ordinal differences in magnitude.

A recent study [Robinson and Darley, 2019] strongly suggests that people are also likely to assign more blame for a repeat violation relative to that assigned for a first-time offense. In the study, subjects were asked to assign hypothetical jail-time penalties for repeated fictional instances of both assault and theft. In both cases, second occurrences incurred more blame than the first. For example, an average penalty of 6.17 months was assigned for the first occurrence of an assault, 6.41 months were assigned for a second assault occurring two weeks later, and 6.59 months were assigned for an assault occurring two years after the initial occurrence. Although the results do not directly account for the influence of perpetrators' mental states on the penalties, the results capture the intuitive notion that people blame more strongly for repeated norm violations.

Having established the psychological theories of blame underlying the proposed framework, we now turn to the formalization of domains.

3 Action Language $\mathcal{AL}$

For the representation of domains and their evolution over time we rely on action language $\mathcal{AL}$ [Baral and Gelfond, 2000]. $\mathcal{AL}$ is centered around a discrete-state-based representation of the evolution of a domain in response to actions. *Fluents* are boolean properties of the domain whose value can change over time, and a *(fluent) literal* is a fluent f or its negation $\neg f$. A *state* σ is a collection of literals, and only an *action* α taken in σ can cause the value of fluents to change in the following state σ'. *Dynamic (causal) laws* describe under what conditions actions change the values of fluents:

$$\alpha \textbf{ causes } l_0 \textbf{ if } l_1, \ldots, l_n$$

Static laws describe relationships between the truth values of fluents in a given state:

$$l_0 \textbf{ if } l_1, \ldots, l_n$$

Finally, *executability conditions* describe under which condition actions cannot occur:

$$\alpha \textbf{ impossible\_if } l_1, \ldots, l_n$$

A set of $\mathcal{AL}$ statements is called an *action description*. The semantics of an action description AD is given by its *transition diagram* $\tau(AD)$, a directed graph $\langle N, A\rangle$ where N is the set of all possible states σ of AD and A is the set of all possible *state transitions* $\langle \sigma, \alpha, \sigma' \rangle$ in accordance with the statements of AD. Finally, a sequence $\langle \sigma_1, \alpha_1, \sigma_2, \ldots, \alpha_k, \sigma_{k+1} \rangle$ is a *path of* $\tau(AD)$ of length k if every $\langle \sigma_i,\ \alpha_i,\ \sigma_{i+1} \rangle$ is a transition in $\tau(AD)$.

4 Norm-Guided Blaming Framework

In this section, we leverage insights from the cited psychological studies to define a framework that aims to facilitate norm-guided blaming in evolving domains. Building upon the constructs of $\mathcal{AL}$, the framework consists of high-level characterizations of norms and violations, the mental states and violation histories of agents, and the effects of both on the amount of blame agents receive for their actions.

Running example. We use Scenario 1 as a running example. Action description AD_e describes the scenario's domain and contains the following dynamic law:

$$\begin{aligned} &worksOn(A, C) \textbf{ causes } shorted(C) \\ &\qquad\qquad \textbf{if } ungrounded(T), tools(T), \\ &\qquad\qquad\quad using(T, A), agent(A), \\ &\qquad\qquad\quad computer(C) \end{aligned}$$

This law states that an agent working on a computer with ungrounded tools will cause a short circuit. AD_e also contains the following static law,

$$damaged(C) \textbf{ if } shorted(C), computer(C)$$

which states that a short-circuited computer is also a damaged computer. Let $\rho_e = \langle \sigma_1, \alpha_1, \sigma_2 \rangle$ be a path in $\tau(AD_e)$ representing the scenario in which Neil damages Allen's computer. The initial state σ_1 contains (at least) literals defining agents in the scene ($agent(neil)$ and $agent(allen)$), information about the tools that Neil is using ($tools(t)$, $using(t, neil)$, and $ungrounded(t)$), and, finally, information about Allen's computer ($computer(c)$, $\neg shorted(c)$, $\neg damaged(c)$, and $propertyOf(c, allen)$). When action $\alpha_1 = worksOn(neil, c)$ occurs in σ_1, both $shorted(c)$ and $damaged(c)$ become true in the resulting state σ_2.

Norms, violations, and causing agents. We now present some preliminary notions that will aid in our discussion of characterizing the effects of mental states and repeated violations on assigning blame. In our framework, a *norm* is represented by a tuple $\eta = \langle C, o \rangle$, which describes the conditions (given by the set of literals C), under which the outcome o of any agent action α is impermissible. We characterize the property damage norm from our example as follows:

$$\eta_e = \langle \{agent(A), \neg propertyOf(P, A)\}, damaged(P) \rangle$$

This norm states that it is not permissible for any action taken by an agent A that results in damage to another agent's property. We denote a set of norms by $N = \{\eta_1, \ldots, \eta_n\}$. A *norm violation* occurs when some action brings about a norm violating outcome o. While it is outside of the scope of this paper to characterize causation for $\mathcal{AL}$, recent research has defined notions of both direct and indirect orthodox causation (i.e., non-omissive) for the language [LeBlanc *et al.*, 2019] which can be integrated into this framework in a straightforward way. In lieu of incorporating these (or related) causal constructs at this stage of the research, we say that an action α *violates* norm η if it occurs while all literals in C are members of σ and α causes (by some future definition) the norm-violating outcome o to "appear" in the resulting state σ'. Similarly, we say that agent A is a *violating agent* of norm η if it performed an action α that violates η. All norm violations are associated with an initial amount of blame, $b(\eta)$, which can be determined by any number of factors (e.g., perceived severity of the violation without knowledge of the norm violating agent or their mental states). We denote norm violations by $v = \langle \eta, \alpha, b(\eta) \rangle$.

For our example, it is straightforward to see that when Neil shorted computer c by working on it with ungrounded tools (AD_e's dynamic law), an "indirect" effect of this action was that the computer was damaged (AD_e's static law). Because our norm η_e states that it is not permissible to perform any action that results in damage to another's property, we conclude that Neil's action of working on the computer with ungrounded tools violated η_e and so Neil is a violating agent of that norm. We represent the example violation as $v_e = \langle \eta_e, workedOn(neil, C), b(\eta_e) \rangle$.

Mental states, histories of violations, and blame intensity. We now present an approach to representing the effects of an agent's mental states and their history of norm-violating behavior on the amount of blame that they receive for violating a norm η. An agent is represented by a tuple $A = \langle M, \gamma \rangle$. M is a set of fluents denoting the agent's mental states[2], where the value of $intentional$ depends on whether or not the agent acted intentionally, $justifiable$ is true if the agent's reasons for acting were good and false otherwise, and $preventable$ is true when the agent was capable/obligated to prevent the violation and false otherwise. If a piece of mental state information is unavailable, the associated literal is omitted from M. An agent's γ is a set of A's previous norm violations. In our example, we represent Neil's mental state and violation history as $agent(neil) = \{M, \gamma\}$, where M contains the literals $\{\neg intentional, preventable\}$, because he could have prevented the outcome by grounding his tools, and $\gamma = \{v_e\}$ because this is his first violation.

We define changes in blame intensity for our framework as follows.

Definition 1. *Given a norm violation $v = \langle \eta, \alpha_i, b(\eta) \rangle$ and the violating agent $A = \langle M, \gamma \rangle$, the* change in blame Δ *relative to $b(\eta)$ is given by the tuple $\langle \delta_{mental}, \delta_{previous} \rangle$, where*

$$\delta_{mental} = \begin{cases} \uparrow\uparrow & M = \{intentional\} \\ \uparrow\uparrow\uparrow & M = \{intentional, \neg justifiable\} \\ \downarrow\downarrow\downarrow & M = \{intentional, justifiable\} \\ \downarrow\downarrow & M = \{\neg intentional\} \\ \downarrow & M = \{\neg intentional, preventable\} \\ \downarrow\downarrow\downarrow & M = \{\neg intentional, \neg preventable\} \end{cases}$$

and $\delta_{previous} = n$, where n is the number of violations of norm η in γ prior to α_i.

The cases of δ_{mental} in Definition 1 capture how a causing agent's mental states influence how much blame is assigned according to the Socially Regulated Blame Perspective. As in their work, arrows indicate magnitudes of blame change(↑ for increase; ↓ for decrease) and the number of arrows indicates ordinal differences in magnitude. The value of $\delta_{previous}$ accounts for the number of previous violations. In our example, the change in blame relative to $b(\eta_e)$ is $\Delta_e = \langle \downarrow\downarrow, \downarrow \rangle$ because Neil's violation was unintentional and preventable. If Neil had violated this norm once before, however, then γ would contain a violation tuple and $\delta_{previous}$ would equal 1.

5 Conclusion

We have proposed a high-level framework for modeling the dynamics of norm-guided blaming that aims to unite psychological theories about the relationships between norms, violations, and the mental states and behaviors of agents. Additional human subject studies are needed to learn how mental state info plays into blaming for repeated violations. Key insights from the literature [Chockler and Halpern, 2004; Tomai and Forbus, 2007; Mao and Gratch, 2005] will aid us in advancement of causal representation and reasoning techniques for the framework.

[2] As with causation, the mental state constructions are placeholders for more sophisticated techniques to be integrated as the development of the framework continues.

References

[Balduccini and Gelfond, 2003] Marcello Balduccini and Michael Gelfond. Diagnostic reasoning with a-prolog. *arXiv preprint cs/0312040*, 2003.

[Baral and Gelfond, 2000] Chitta Baral and Michael Gelfond. Reasoning agents in dynamic domains. In *Logic-based artificial intelligence*, pages 257–279. Springer, 2000.

[Chockler and Halpern, 2004] Hana Chockler and Joseph Y Halpern. Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, 22:93–115, 2004.

[LeBlanc *et al.*, 2019] Emily LeBlanc, Marcello Balduccini, and Joost Vennekens. Explaining actual causation via reasoning about actions and change. In *European Conference on Logics in Artificial Intelligence*, pages 231–246. Springer, 2019.

[Malle *et al.*, 2014] Bertram F Malle, Steve Guglielmo, and Andrew E Monroe. A theory of blame. *Psychological Inquiry*, 25(2):147–186, 2014.

[Mao and Gratch, 2005] Wenji Mao and Jonathan Gratch. Social causality and responsibility: Modeling and evaluation. In *International Workshop on Intelligent Virtual Agents*, pages 191–204. Springer, 2005.

[Monroe and Malle, 2019] Andrew E Monroe and Bertram F Malle. People systematically update moral judgments of blame. *Journal of Personality and Social Psychology*, 116(2):215, 2019.

[Robinson and Darley, 2019] Paul H Robinson and John M Darley. *Justice, liability, and blame: Community views and the criminal law (Study 18)*. Routledge, 2019.

[Tomai and Forbus, 2007] Emmett Tomai and Ken Forbus. Plenty of blame to go around: a qualitative approach to attribution of moral responsibility. Technical report, Northwester Univ Evanston, IL, 2007.