Transgression and Atonement

Kevin M. Knight and **Deepthi Chandrasekaran** and **Aline Normoyle**
Ransom Weaver and **Barry G. Silverman**
Electrical and Systems Engineering Department
University of Pennsylvania, Philadelphia, PA 19104-6315

Abstract

This paper presents an approach to modeling social transgressions in agent based systems. The approach is intended to be abstract enough that it may be used with many different theories of transgression, apology, forgiveness, etc. We discuss an implementation of this approach in PMFserv, an agent based socio-cognitive modeling framework.

Introduction

In this paper we will present an approach to modeling social transgressions in agent based systems. By "social transgression" we mean an offense an agent can commit against social rules. Throughout this paper, the terms "transgression" and "offense" (and, similarly, "transgressor" and "offender") will be used interchangeably.

Our approach involves representing transgressions as abstract objects. However, the transgression objects themselves are not the main focus. Rather, they simply serve as a nexus for actions and relations between agents about the transgressions they represent.

Transgressions

In this section we will discuss issues with transgressions in general. One notable issue that we will not consider is perceptual mistakes. For example, we will not address situations in which an observer incorrectly blames an innocent party for a transgression or underestimates the effects. Such mistakes are beyond the scope of this paper.

All transgressions have a transgressor, a set of victims, and a set of effects. Effects are to be understood as the direct effects of the offending action, not the emotional effects on observers. Those are handled separately. For example, if Alan sets fire to Brad's house and burns it down, then the transgressor is Alan, the victims are Brad and whoever else has a stake in his house, and the set of effects is that Brad's house has burned down and most everything inside has been damaged or destroyed.

Now, one might argue that a transgression may have multiple transgressors. Take the example of a bank robbery executed by a gang with four members. *Prima facie*, this seems to be just such a transgression. However, we would consider it to be four separate transgressions – one for each robber. Or, in any case, we represent it as four separate transgression objects.

Metaphysical issues aside, we have pragmatic reasons for this method of dividing transgressions. First, it allows us to distinguish the different roles and levels of guilt of the different transgressors. For example, in a bank robbery, the gang member whose only job was to drive the getaway car may be held to a lower level of responsibility than the gang members who threatened the people in the bank with weapons.

Second, it allows us to keep relations between the different transgressors and the transgression separate. As we will discuss below, the transgression objects are used to keep track of such things as whether the transgressor committed the transgression intentionally. Since different transgressors may have had different levels of intent, even in what may be considered the same transgression, we use multiple objects to keep these relations separate.

Indeed, similar arguments can be made on the victim side. In many cases it is reasonable to create a separate transgression for each transgressor-victim pair. This allows for a very detailed level of accounting. However, in other cases such an approach may be infeasible. In the bank robbery example, there is a potentially very large number of victims, including the bank itself, all of its shareholders, everyone in the bank at the time of the robbery, and everyone whose money is lost. Whether to allow multiple victims is something that can be decided on a model by model or even transgression by transgression basis.

Beyond these basic properties of transgressions themselves, our transgression objects will keep track of some relations with the transgressor, relations with observers, properties of the effects, and relations between the transgressor and observers.

A transgression may be intentional or unintentional. For example, Alan may be angry at Brad and intentionally run his car into Brad's car. On the other hand, Alan may run into Brad's car accidentally.[1]

[1]Notice that intention is distinct from responsibility. That Alan ran his car into Brad's unintentionally does not imply that he is not responsible. For example, he might have been negligent in his driving.

The effects of a transgression may be active or inactive. Let us assume in our example of Alan crashing his car into Brad's that Brad's car is damaged and his arm is broken. Until his car is fixed (or replaced) and his arm heals, the effects of the transgression are active. Once those things happen, the effects are inactive.[2]

A related but separate issue is compensation. That is, some or all of the victims may have received compensation for the harm done to them. Compensation may or may not come from the transgressor and may or may not cause effects to become inactive. In our car crash example, compensation will probably come from Alan's insurance company rather than from Alan himself, and it will probably come in the form of money intended to cover repairs to or replacement of Brad's car as well as medical expenses. In this case, the compensation does not make the effects inactive. As noted before, the effects remain active until Brad's car is repaired or replaced and his arm heals.

A transgressor may or may not have apologized for the transgression. Apology is a complex subject, and there is much to say about it, both in terms of structure and effects. For the purposes of this paper, we will consider apology to be a black box. We will touch on the effects when we discuss forgiveness.

A transgression may be forgivable or unforgivable. It seems that most people view most transgressions as, at least in principle, forgivable. However, some people may view some transgressions as unforgivable, at least until some condition occurs (such as repentance of the transgressor).

Among forgivable transgressions, a transgression may be forgiven or unforgiven. This means that the observer in question may have forgiven the transgressor for the transgression. We will discuss forgiveness further in a later section.

Emotional Reactions

Any observer could potentially have an emotional reaction to a transgression. This includes direct observers (i.e., those who directly perceive the transgression) and indirect observers (i.e., those who learn about the transgression by means other than direct perception, such as newspapers or other observers). It also includes those who have some relationship to those directly involved and those who have no such relationship.

For any transgression, we should expect that there is someone who would have a negative emotional reaction (e.g., anger or reproach) to observing it. However, this does not imply that everyone would have the same negative emotional reaction. For example, Wunderle points out that "Arabs believe it is imperative that negotiating partners respect each other's honor and dignity. To an American, losing face may be embarrassing, but to an Arab, it is devastating" (Wunderle 2007, p. 36). Even within a culture, there is considerable variation between individuals in the severity of their emotional reactions to the same transgression (see, e.g., (Azar, Mullet, and Vinsonneau 1999; McCullough, Fincham, and Tsang 2003; Wohl and McGrath 2007)).

In addition to cultural and personality factors, the relationship between the observer and those directly involved in the transgression may affect the extent of the reaction. Gordijn, Wigboldus, and Yzerbyt (2001) and Yzerbyt et al. (2003) studied emotional reactions of uninvolved observers to transgressions. Both found that negative emotional reactions to transgressions are significantly stronger when the victims are in the same group as the observer.

For many transgressions, we should expect that there is someone who would not view it as a transgression. For example, killing cattle is commonplace in America but taboo in India. Indeed, what is an egregious transgression to one may be a cause for celebration to another. Consider Bobby Fischer's reaction to the September 11, 2001 World Trade Center attack. During a radio interview in the Philippines hours after the event, he is reported as describing news of the attack as "wonderful" and saying that he "applaud[s] the act" (Bamber and Hastings 2001).

In a nutshell, our framework must accommodate a wide variety of reactions to a transgression. In particular, it must handle different individuals viewing the same transgression as having different degrees of severity, as well as individuals who do not view the act as a transgression at all. However, since it is a framework for transgressions (and not acts in general), it need not handle the emotional reactions of those who do not view the act as a transgression (though it must not force a negative reaction upon them).

Forgiveness

We will consider what is sometimes called "offense-specific" forgiveness. This is a relationship between three entities: a forgiver, a transgressor, and a transgression. The forgiver forgives the transgressor for the transgression. We are not concerned with whether or how the forgiver is connected to the transgressor or the transgression. However, the forgiver must be aware of the transgression and believe that the transgressor is in some way responsible for it.

We will divide forgiveness along two axes. The first axis is active versus passive forgiveness. *Active forgiveness* is where someone has made a conscious decision to forgive a transgressor; *passive forgiveness* is where no conscious decision has been made.[3]

The second axis is effective versus ineffective forgiveness. *Effective forgiveness* is where the negative emotions toward the transgressor resulting from the transgression have subsided; *ineffective forgiveness* is where the negative emotions have not subsided.

There are three possible combinations of these: effective active forgiveness, ineffective active forgiveness, and effec-

[2] It is worth noting that not all effects can be made inactive. For example, in the case of a murder, the death may never be undone. However, the case of a transgression with permanent effects should not be confused with the case of an unforgivable transgression. Many transgressions with permanent effects may still be forgiven.

[3] Active forgiveness should be understood as a private decision. Whether or not that decision is communicated to anyone else is a separate issue.

tive passive forgiveness. Ineffective passive forgiveness is not really forgiveness, since neither the intent to forgive nor the desired result of forgiving is present.

Many (and perhaps most) definitions of forgiveness reflect a type of effective forgiveness.

- Subkoviak et al. define forgiveness as the "absence of negative affect, judgment, and behavior toward an offender and the presence of positive affect, judgment, and behavior toward the same offender" (1995, p. 642).
- McCullough, Worthington, and Rachal define forgiveness as "the set of motivational changes whereby one becomes (a) decreasingly motivated to retaliate against an offending relationship partner, (b) decreasingly motivated to maintain estrangement from the offender, and (c) increasingly motivated by conciliation and goodwill for the offender, despite the offender's hurtful actions" (1997, pp. 321-322).
- Berry et al. define forgiveness as "the juxtaposition or superimposition of strong, positive, other-oriented emotions over the negative emotions of unforgiveness" (2005, p. 186).

Wohl, Kuiken, and Noels (2006) caution against definitionally rejecting active ineffective forgiveness, describing it as *failed* forgiveness rather than non- or pseudo-forgiveness.[4] We will follow their lead on this point.

Three things are notable about the above definitions. First, none requires a conscious decision to forgive; thus, all are consistent with both active and passive forgiveness. Second, only the first definition requires a behavioral change (though we may expect behavioral changes to accompany the motivational or emotional changes required by the other two). Third, all involve both a decrease in negative emotions and an increase in positive emotions.

Regarding the third point, there is some evidence that the decrease in negative emotions is a separate process from the increase in positive emotions (McCullough, Fincham, and Tsang 2003). This is why an increase in positive emotions is not included in the definition of effective forgiveness.

Now let us consider the three cases of forgiveness, beginning with passive effective forgiveness. In this case, there has been no conscious decision to forgive but the negative emotions resulting from the transgression have subsided. One might expect that this state will occur with time, and there is evidence that this is correct.

McCullough, Fincham, and Tsang (2003) showed that negative emotions (specifically avoidance and revenge motivation) associated with a transgression decrease linearly over time.[5] Moreover, while the rate of decrease varies from person to person, it does not depend on the severity of the transgression. (On the other hand, the initial intensities of the negative emotions caused by the transgression do depend on its severity.) Importantly, they provide not only a theory of *whether* forgiveness will occur but also *when*.

Thus our framework must handle emotion decay, at least regarding the emotions caused by transgressions. However, while McCullough, Fincham, and Tsang have suggested that the rate of decay is linear and does not depend on the severity of the transgression, these assumptions are not built into the framework.

Next let us consider active effective forgiveness. This involves both a conscious decision to forgive and subsidence of negative emotions. Azar, Mullet, and Vinsonneau (1999) studied the effects of four factors on the propensity to forgive. The four factors were (1) whether the transgressor apologized, (2) whether the effects were still active, (3) whether the transgression was intentional, and (4) whether the transgressor was in the same social group as the potential forgiver. They found that the first two had major (and roughly equal) effects, the third a moderate effect, and the fourth an insignificant effect. Moreover, the effects combined additively.

Our transgression objects make available all the information that Azar, Mullet, and Vinsonneau designated as pertinent, including the social relationship between the observer and transgressor. Unfortunately, while they provide insight into how these factors affect whether forgiveness will occur, they provide no insight into how the factors affect when it will occur. Nonetheless, our framework must be able to accommodate different theories about how these (or, indeed, other) factors affect both whether and when forgiveness will occur.

Finally, we will consider active ineffective forgiveness. This involves a conscious decision to forgive, but little or no subsidence of the negative emotions resulting from the transgression in question. This case should not be confused with the case in which the negative emotions caused by the transgression subside but are replaced by further transgressions.

Wohl, Kuiken, and Noels (2006) refer to this case as failed forgiveness. In their study on different types of forgiveness, they identified a type in which the forgiver attempted to resume a positive relationship with the transgressor without ignoring or forgetting the transgression. In such cases, the relationships between forgiver and transgressor tended to deteriorate in the long run.

Unfortunately, we have very little insight into how or why such failed forgiveness might occur or what its precise effects are, including how and under what circumstances the relationship might deteriorate.

[4] They say that for their subjects, "this profile of activities constitutes forgiveness even though forgiveness – as they conceive it – has failed to achieve the desired consequences" (Wohl, Kuiken, and Noels 2006, p. 558).

[5] Wohl and McGrath (2007) confirmed these results and further noted that it is the *perceived* rather than actual amount of time that has passed that affects forgiveness. That is, avoidance and revenge motivation decrease with increases in perceived temporal distance. Since perceived temporal distance fluctuates, so do avoidance and revenge motivation.

Goals	Standards	Preferences
Belonging	Conformance to society	Humanistic
Esteem	Relationship vs. task focus	Materialistic
Safety	Sensitivity to life	Symbolistic
	Use of violence	
	Honesty	
	Respect for authority	
	Self-interest vs. greater good	

Table 1: Example goals, standards, and preferences

PMFserv

PMFserv (Performance Moderator Function Server) is a framework for modeling socio-cognitive agents. It includes a synthesis of about 100 best-of-breed models of personality, culture, values, emotions, stress, social relations, and group dynamics, as well as an integrated development environment for authoring and managing reusable archetypes and their task sets. For each agent, PMFserv operates its perception, physiology, personality, and value system to determine stressors, grievances, tension buildup, the impact of speech acts, emotions, and various mobilization and collective and individual action decisions. PMFserv also manages social relationship parameters and thus macro-behavior (e.g., in collectives or crowds of agents) emerges from individuals interactions and micro-decisions.

PMFserv is in use by an intelligence agency to model diplomatic decisions of world leaders for which it has passed statistical correspondence tests showing it is significantly in agreement with their decision making (Silverman et al. 2007; 2008). PMFserv has also reached the level where it can realistically simulate ethno-political conflicts among regional leaders and their followers vying over control of contested resources and assets. For more detailed accounts of PMFserv, including validation studies for application in the Far East, Middle East, Africa, and North America, see (Silverman et al. 2006b; 2006a; 2007; 2008).

Goals, Standards, and Preferences

Agents' cultural values and personality traits are modeled in PMFserv by goals, standards, and preference (GSP) trees. These are multi-attribute value structures where each tree node is weighted with Bayesian importance weights.

Preferences are long term desires for world situations and relations. In the implementation we describe below, relevant preferences include whether the agent has a materialistic, symbolistic, or humanistic vision of the future.

Goals cover short-term needs and motivations that implement progress toward preferences. Goals relevant to the implementation we describe below include needs for belonging, esteem, and safety.

Standards define the methods an agent is willing to use to satisfy its goals and preferences. These include concerns with conformance to society, relationship versus task focus, sensitivity to life, willingness to use violence, concern with honesty, respect for authority, and narrow self-interest versus concern for the greater good.

The example goals, standards, and preferences just mentioned are summarized in Table 1. It should be noted that these are just examples which are relevant to this article and by no means exhaust the set of possible goals, standards, and preferences.

In addition to Bayesian importance weights, the nodes in the GSP trees have positive and negative activations. A node becomes activated when an agent takes an action related to that node. For example, if an agent takes an action involving deceit, then the node representing its standard of honesty would be negatively activated. These activations are used to calculate the agent's current emotional state.

GSP trees and how they relate to emotions in PMFserv have been discussed at length elsewhere (see, e.g., (Silverman et al. 2006b; 2007)), and we will not reproduce that discussion here.

Objects

In addition to managing agents, PMFserv manages objects (representing both agents and non-agents, such as a car or a location), including when and how they may be perceived and acted on by agents. PMFserv implements affordance theory, meaning that each object applies perception rules to determine how it should be perceived by each perceiving agent. Objects then reveal the actions (and the potential results of performing those actions) afforded to the agent. For example, an object representing a car might afford a driving action which can result in moving from one location to another.

Notably, objects need not be concrete. PMFserv makes no metaphysical assumptions about its objects. Abstract objects, such as plans and obligations, may be represented just as easily as concrete objects.

Objects have a state, which is a set of properties. For example, an object representing a car might have a make, model, color, sale price, etc. Additionally, objects have a set of perceptual types. Each perceptual type has a perceptual rule associated with it which is used to determine whether that type is perceived. For example, a car might have a *buyable* perceptual type which indicates whether an agent perceives the car as something it can purchase. The perceptual rule associated with the type might compare the sale price of the car with the amount of money the agent has (as well as considering whether the car is owned by someone else and is for sale). Assuming that the agent perceives the car as *buyable*, the action *buy* would be afforded with the result that the car changes ownership, the current owner's money increases, the agent's money decreases, and the agent's emotional state changes appropriately.

In addition to binary perceptual types, there are "continuous" perceptual types. Rather than an agent viewing an object as either having this sort of perceptual type or not, agents view an object as having it to a certain degree between 0 and 1. The degree of perception and the precise meaning of the degree are determined by perceptual rules. For example, an agent might view a glass of water as more or less full or another agent as more or less of a friend.

Furthermore, groups of perceptual types for an object may be designated as mutually exclusive. That is, at most one of such a group may be perceived at a time by an agent. A perceptual type may be in at most one such group for an object.

Emotion Decay

In PMFserv, whenever an event occurs which should elicit an emotional reaction from an agent, the agent notes the event along with its initial emotional impact and assigns a decay function to it. An agent's emotional state at any given time is determined by the initial impact, decay function, and age of each event that it has stored.

Each decay function takes the initial impact and age of an event and returns the decayed impact, i.e., the impact the event will have after a certain amount of time has passed. In principle, there is no limitation to the nature of the function, though under normal circumstances it should be monotonically decreasing.

Each agent has its own decay policy which assigns decay functions to events. Like the decay functions, there are no real limits to their nature. A decay policy could, for example, assign the same decay function to all events; or it could assign decay functions based on properties of the events.

Implementation

In PMFserv, transgressions are represented as abstract objects. They are dynamically created when an agent transgresses. This requires an account of what actions count as transgressions and what impact they will have on observers. These may vary significantly between scenarios since they depend on the actions that are available and the sorts of agents being modeled.

In our current implementation, we are modeling Arab villagers and US soldiers in an Iraqi village. Emphasis in this article is on transgressions that US soldiers can commit against villagers and how they may atone (though the implementation also handles transgressions between villagers). While there are in fact a vast number of such transgressions, for this discussion we will concentrate on three examples: rude and untactful speech (*adeb*), searching a home without dogs, and searching a home with dogs. Both cases of searching refer to soldiers searching a villager's home by force or the threat of force.

Before discussing the transgressions any further, we will discuss some simplifying assumptions we are making, mostly with respect to perception and communication.

The first assumption is that all transgressions are perceived immediately by everyone. This does not necessarily mean that everyone directly perceives every transgression, simply that everyone is immediately aware. Essentially, we are assuming that communication about transgressions within the village is complete and effectively instantaneous.

The second assumption is that transgression objects have only one victim. Thus transgressions which have multiple victims will be represented by multiple transgression objects each with one victim.

The third assumption is that only Arab villagers are offended. That is, we are not representing transgressions that villagers can commit against soldiers or soldiers against each other.

Now let us put this into a more formal representation. Let T be the set of transgressions, A be the set of agents, E be the set of effects, and $\langle TP, \prec, d\rangle$ be a structure representing time, where TP is the set of timepoints, $\prec$ is a linear ordering, and d is a distance function. We will represent a transgression $\tau \in T$ as a quadruple $\langle o, v, e, t\rangle$, where $o \in A$ is the offender (or transgressor), $v \in A$ is the victim, $e \subseteq E$ is the set of effects, and $t \in TP$ is the time at which the transgression occurred. For a transgression τ, we will denote these as τ_o, τ_v, τ_e, and τ_t, respectively.

Based on earlier discussion we will define the following predicates. For $\tau \in T$, $\alpha \in A$, and $t_0, t_1 \in TP$,

- $intentional(\tau)$ is true iff τ was intentional,
- $apologized(\tau, t_1)$ is true iff τ_o apologized for τ at some time $t_0 \preceq t_1$,
- $active(\tau, t_1)$ is true iff τ_e are still active at t_1, and
- $forgivable(\tau, \alpha)$ is true iff observer α views τ as forgivable.

We will define some more functions and predicates after further discussion.

Now that we have stated what our transgressions are, we must say what their impact on observers will be. That is, we must associate each transgression with a set of GSP activations that will be afforded to observers. We can represent afforded activations as a vector in $[0, 1]^{2n}$, where n is the number of GSP nodes. (The vector is of length $2n$ because it must contain both positive and negative activations for each node.)

To facilitate combination of such vectors, we define the bounded addition operator, written $\oplus$. For $x, y \in \mathbb{R}$, we define scalar bounded addition as follows.

$$x \oplus y = \max(0, \min(1, x + y)) \tag{1}$$

We define vector bounded addition as element-wise scalar bounded addition.[6]

As a convenient way to organize these in our implementation, each transgression is assigned an intensity in each of the following categories: *faux pas*, *taboo*, *violent*, *materialistic*, and *deceitful*.[7] Intensities range from zero to one, and it is common for transgressions to have non-zero intensities in multiple categories. For example, a mugging is both violent and materialistic. The meaning of each category is summarized in Table 2.

[6] Scalar bounded addition is commutative in general and associative for non-negative values. Zero is a bounded additive identity for values in $[0, 1]$. Vector bounded addition has analogous properties.

[7] These categories should not be taken as an authoritative taxonomy of transgressions. They were chosen because they correspond well to nodes in the GSP trees used in the current scenario. For research on moral categories, see, e.g., (Haidt 2007).

Faux pas	Taboo	Violent	Materialistic	Deceitful
Relationship focus	Relationship focus	Sensitivity to life	Self-interest	Honesty
Conformance to society	Conformance to society	Use of violence	Respect for authority	
Belonging	Belonging	Belonging	Materialistic	
Esteem	Esteem	Safety		
	Symbolistic			

Table 2: Transgression categories

Faux pas are comparatively minor transgressions related to etiquette. Examples include rude speech and inappropriate dress. These afford activations to an observer's GSP nodes related to focusing on relationships, conformance to society, and concerns with belonging and esteem. Let $fp \in [0,1]^{2n}$ be the activations afforded by a faux pas transgression.

Taboo transgressions are similar in nature to faux pas, though they are generally more serious. Examples include marrying a sibling and making blasphemous statements. These afford activations to the same GSP nodes as faux pas plus those related to symbolistic concerns. Let $tb \in [0,1]^{2n}$ be the activations afforded by a taboo transgression.[8]

Violent transgressions can range from the relatively minor to the extremely serious. Both actual violence and the threat of violence are included. Examples include slapping someone in the face and setting off a bomb in a crowded marketplace. These afford activations to an observer's GSP nodes related to sensitivity to human life, the use of violence, and concerns with belonging and safety. Let $vi \in [0,1]^{2n}$ be the activations afforded by a violent transgression.

Materialistic transgressions are those having to do with property. This includes damaging, destroying, and stealing property. Examples include vandalism and theft. These afford activations to an observer's GSP nodes related to self-interest, respect for authority, and materialistic concerns. Let $ma \in [0,1]^{2n}$ be the activations afforded by a materialistic transgression.

Deceitful transgressions are those relating to honesty. They include everything from little white lies to major fraud. These afford activations to an observer's GSP nodes related to honesty. Let $de \in [0,1]^{2n}$ be the activations afforded by a deceitful transgression.[9]

Intensities for our example transgressions can be found in Table 3. *Adeb* is a relatively minor faux pas. Searching a home (with or without dogs) involves the threat of violence and offense against property. Searching a home with dogs also involves elements of taboo, since dogs are considered unclean by many Arabs.

To denote the intensity of a transgression in each category, we will define five functions from T to $[0,1]$: $fauxpas$, $taboo$, $violent$, $materialistic$, and $deceitful$. The base impact a transgression $\tau \in T$ will have on an observer is defined by the following equation.

$$\begin{aligned} I_b(\tau) \quad = \quad & fauxpas(\tau) \cdot fp \oplus taboo(\tau) \cdot tb \oplus \\ & \oplus violent(\tau) \cdot vi \oplus deceitful(\tau) \cdot de \oplus \\ & \oplus materialistic(\tau) \cdot ma \end{aligned} \tag{2}$$

The initial impact is affected by two other factors: the relationship of the observer to the victim and whether the transgression was intentional.

As noted earlier, the relationship of the observer to the victim can affect the impact of a transgression. In particular, the closer the relationship, the more severe the impact. We consider four types of relationships: whether the observer is the victim, the victim's kin, in the same group as the victim, or in a group with at least neutral relations with the victim's group. If the observer does not share one of these relationships with the victim, then there will be no relationship based impact. For $\tau \in T$ and $\alpha \in A$, we will denote the impact of α's relationship to τ_v by $I_r(\tau, \alpha) \in [0,1]^{2n}$.

Based on studies by (Azar, Mullet, and Vinsonneau 1999), we give additional initial impact if the transgression was intentional. For $\tau \in T$, we denote this impact by $I_n(\tau) \in [0,1]^{2n}$, where $I_n(\tau) = \langle 0, \ldots, 0 \rangle$ if $intentional(\tau)$ is not true.

Thus for $\tau \in T$ and $\alpha \in A$, the initial impact I of α observing τ is described by the following equation.

$$I_i(\tau, \alpha) = I_b(\tau) \oplus I_r(\tau, \alpha) \oplus I_n(\tau) \tag{3}$$

Of course, the actual emotional effect τ will have on α is a function of $I_i(\tau, \alpha)$, α's personality, and α's prior emotional state.

For example, *adeb* will not bother someone who is not concerned with relationships, conformance to society, belonging, or esteem. On the other hand, someone who is concerned with one or more of those will likely be bothered, though probably not too much since *adeb* is a minor transgression at worst.

All of these factors are implemented as perceptual types on the transgression object. The categories are implemented as continuous perceptual types (where the perception levels are the intensities from Table 3), the relationship is implemented as a mutually exclusive group of binary perceptual types, and the intentionality is represented as a single binary perceptual type. These perceptual types afford *perceive* actions, which are performed automatically when the object is introduced.

[8] Faux pas and taboo are notably similar categories. The main difference is that taboo transgressions violate deeply held convictions. While a faux pas transgression might result in feelings of annoyance or perhaps even mild contempt, a taboo transgression would more likely result in feelings of anger or even disgust. Consider the difference between addressing someone in an inappropriate way and throwing feces at that person.

[9] This category is currently a placeholder. At this time, agents are not able to take deceptive actions in PMFserv.

	Faux pas	Taboo	Violent	Materialistic	Deceitful
Adeb	0.1	0.0	0.0	0.0	0.0
Searching a home	0.1	0.0	0.2	0.1	0.0
Searching a home with dogs	0.1	0.5	0.5	0.1	0.0

Table 3: Example transgression intensities

There are two binary perceptual types on the transgression object which afford substantive actions. The first is perceivable if the effects of the transgression are still active and affords the action *remove effects*. The second is perceivable only to the transgressor if he has not apologized and affords the action *apologize*.

The two actions are similar in effect. Both reduce the emotional impact of the transgression on observers, thus decreasing the time it takes to forgive. This is based on the claim of (Azar, Mullet, and Vinsonneau 1999) that whether the effects of the transgression are still active and whether the transgressor has apologized significantly contribute to the likelihood of forgiveness. Notably, neither action can be performed multiple times for the same transgression.

The *apologize* action has an added dimension in our model since some transgressions require atonement more complicated than a simple verbal apology. Consider, for example, the ritual of "blood money" paid for an offense resulting in death in Arab cultures.[10] To this end, we include atonement objects, which encapsulate the steps necessary for atonement. Once all the steps have been completed, the effect is that of having apologized.

Formally, performing the *remove effects* or *apologize* on transgression $\tau \in T$ at time $t \in TP$ has the effect of making $active(\tau, t^{'})$ false or $apologized(\tau, t^{'})$ true for all $t^{'} \in TP$ such that $t \preceq t^{'}$. We implement the reduction in emotional impact by associating coefficients with each as follows.

$$C_e(\tau, t) = \begin{cases} 0 & \text{if } active(\tau, t) \\ -\frac{1}{2} & \text{otherwise} \end{cases} \quad (4)$$

$$C_a(\tau, t) = \begin{cases} -\frac{1}{3} & \text{if } apologized(\tau, t) \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

These are used to adjust the impact as follows.

$$I(\tau, \alpha, t) = (1 + C_e(\tau, t) + C_a(\tau, t)) \cdot I_i(\tau, \alpha) \quad (6)$$

In other words, removing the effects reduces the impact by half and apologizing reduces it by one third.[11]

Furthermore, there is the question of how the impact decays over time. Each agent is assigned a "grudge factor" ranging from 0 to 1 and indicating for how long the agent will hold a grudge. The higher the grudge factor, the longer it will take the agent to forgive a transgression. In practical terms, this determines the emotion decay function for that agent. Following (McCullough, Fincham, and Tsang 2003), all decay functions are linear, and the grudge factor simply serves to determine the slope (with a lower grudge factor indicating a steeper slope).[12] Slopes range from -1 (indicating more or less instantaneous forgiveness) to 0 (indicating no forgiveness).

The only exception is for unforgivable transgressions. For those transgressions, emotions do not decay. Such transgressions are rare, and none of our examples fall into this category. Whether a transgression is unforgivable is implemented as a binary perceptual type on the transgression object.

For $\alpha \in A$, let us denote the slope of α's decay function by $\alpha_d \in [-1, 0]$. Thus for $\tau \in T$ and $t \in TP$ such that $\tau_t \preceq t$, the amount that the impact of τ should have decayed by t is described by the following equation.

$$\delta(\tau, \alpha, t) = \begin{cases} \alpha_d \cdot d(t, \tau_t) & \text{if } forgivable(\tau, \alpha) \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

(where d is a temporal distance function). Now we define the decayed impact of τ on α at t as

$$D(\tau, \alpha, t) = I(\tau, \alpha, t) \oplus \vec{\delta}(\tau, \alpha, t) \quad (8)$$

where $\vec{\delta}(\tau, \alpha, t)$ is a vector in $\mathbb{R}^{2n}$ whose elements are all $\delta(\tau, \alpha, t)$.

Once the emotional impact of a transgression has decayed to nothing, then the agent has effectively forgiven the transgression. In other words, for $\tau \in T$, $\alpha \in A$, and $t \in TP$ such that $\tau_t \preceq t$, α has effectively forgiven τ_o for τ at t if $D(\tau, \alpha, t) = \langle 0, \ldots, 0 \rangle$.

Consider a few examples with the sample transgressions mentioned earlier. If a soldier commits *adeb*, most villagers will have a slightly negative emotional reaction. However,

[10] The ritual may be fairly elaborate as, for example, described by (Irani and Funk 1998). In this case, the family of the offender must seek the help of a delegation of local leaders, esteemed mediators, and other notables, who will hear the grievances of the victim's family and determine what constitutes an appropriate payment of "blood money" in the case at hand. Then the offending and offended families gather together for a ritual shaking of hands. Then the family of the victim serves bitter coffee to the family of the offender to demonstrate forgiveness. Finally, the offending family serves a meal to the offended family.

[11] The actual values of these coefficients are not supported by the literature. They are initial guesses on our part. However, their independence is supported by (Azar, Mullet, and Vinsonneau 1999).

[12] Our assignment of grudge factors, and thus decay rates, to agents is somewhat arbitrary. We made what we consider to be plausible guesses, but as far as we can tell, the literature is largely silent on this issue.

for the most part they will get over it quickly, especially if the soldier apologizes.

Searching a home is a more serious transgression, involving violent and materialistic elements as well as breaching etiquette. Villagers will have a much stronger reaction than to *adeb*. However, once the effects become inactive and the soldier apologizes, most villagers should forgive the soldier for that particular transgression within a few weeks (though the villagers may still have negative emotions about the soldier if he has committed further transgressions).

Searching a home with a dog is a considerably more severe transgression than either of the previous two. In addition to the effects of simply searching a home, this violates the taboo of bringing a dog into a home. Thus the emotional impact on the villagers will be considerably stronger. Even once the effects have been removed and the soldier has apologized, forgiveness may take quite some time, perhaps several months (with the same qualification as before). And without an apology forgiveness will take considerably longer.

Conclusion

We have presented an approach to modeling transgressions in agent based systems. To this end we have discussed a number of considerations relevant to any model of emotional reaction to and forgiveness of transgressions. And we have described an implementation in PMFserv.

There are still many open issues on this topic. We did not consider the question of observers having incomplete or incorrect information about a transgression. Similarly, there are issues we did not consider with communication, such as agents (intentionally or unintentionally) introducing their own biases when informing others of a transgression. In the real world these are very common cases.

The issue of collective responsibility remains open. That is, how observers attribute blame to groups for individual transgressions. For example, when a US soldier commits a transgression, how much will observers blame the soldier himself versus the US military versus the US as a whole?

Another interesting issue we did not consider is apology. There is a great deal to say on the subject, particularly regarding the effectiveness of different apology strategies and the likelihood of an apology being rejected.

Finally, beyond conceptual issues, for any approach to modeling transgressions to be really useful, actual transgressions and their impacts must be cataloged.

References

Azar, F.; Mullet, E.; and Vinsonneau, G. 1999. The propensity to forgive: Findings from Lebanon. *Journal of Peace Research* 36(2):169–181.

Bamber, D., and Hastings, C. 2001. Bobby Fischer speaks out to applaud Trade Center attacks. *Sunday Telegraph (London)* 17.

Berry, J. W.; Worthington, E. L.; O'Connor, L. E.; Parrott, L.; and Wade, N. G. 2005. Forgivingness, vengeful rumination, and affective traits. *Journal of Personality* 73(1):183–225.

Gordijn, E. H.; Wigboldus, D.; and Yzerbyt, V. 2001. Emotional consequences of categorizing victims of negative outgroup behavior as ingroup or outgroup. *Group Processes & Intergroup Relations* 4(4):317–326.

Haidt, J. 2007. The new synthesis in moral psychology. *Science* 316(5827):998–1002.

Irani, G. E., and Funk, N. C. 1998. Rituals of reconciliation: Arab-Islamic perspectives. *Arab Studies Quarterly* 20(4):53–74.

McCullough, M. E.; Fincham, F. D.; and Tsang, J.-A. 2003. Forgiveness, forbearance, and time: The temporal unfolding of transgression-related interpersonal motivation. *Journal of Personality and Social Psychology* 84(3):540–557.

McCullough, M. E.; Worthington, E. L.; and Rachal, K. C. 1997. Interpersonal forgiving in close relationships. *Journal of Personality and Social Psychology* 73(2):321–336.

Silverman, B. G.; Bharathy, G. K.; O'Brien, K.; and Cornwell, J. B. 2006a. Human behavior models for agents in simulators and games: Part II: Gamebot engineering with PMFserv. *Presence: Teleoperators and Virtual Environments* 15(2):163–185.

Silverman, B. G.; Johns, M.; Cornwell, J. B.; and O'Brien, K. 2006b. Human behavior models for agents in simulators and games: Part I: Enabling science with PMFserv. *Presence: Teleoperators and Virtual Environments* 15(2):139–162.

Silverman, B. G.; Bharathy, G. K.; Nye, B.; and Eidelson, R. J. 2007. Modeling factions for "effects based operations": Part I – leaders and followers. *Computational & Mathematical Organization Theory* 13(4):379–406.

Silverman, B. G.; Bharathy, G. K.; Nye, B.; and Smith, T. 2008. Modeling factions for "effects based operations": Part II – behavioral game theory. *Computational & Mathematical Organization Theory* 14(2):120–155.

Subkoviak, M. J.; Enright, R. D.; Wu, C.-R.; Gassin, E. A.; Freedman, S.; Olson, L. M.; and Sarinopoulos, I. 1995. Measuring interpersonal forgiveness in late adolescence and middle adulthood. *Journal of Adolescence* 18(6):641–655.

Wohl, M. J. A., and McGrath, A. L. 2007. The perception of time heals all wounds: Temporal distance affects willingness to forgive following an interpersonal transgression. *Personality and Social Psychology Bulletin* 33(7):1023–1035.

Wohl, M. J. A.; Kuiken, D.; and Noels, K. A. 2006. Three ways to forgive: A numerically aided phenomenological study. *British Journal of Social Psychology* 45(3):547–561.

Wunderle, W. 2007. How to negotiate in the Middle East. *Military Review* 87(2):33–37.

Yzerbyt, V.; Dumont, M.; Wigboldus, D.; and Gordijn, E. 2003. I feel for us: The impact of categorization and identification on emotions and action tendencies. *British Journal of Social Psychology* 42(4):533–549.