A Computational Approach to Etiquette and Politeness: Initial Test Cases

Dr. Christopher Miller
Ms. Peggy Wu
Mr. Harry Funk
Dr. Peggy Wilson
Smart Information Flow Technologies
211 First St. N. Ste. 300
Minneapolis, MN 55401
(612) 339-7438, (612) 339-7437
{cmiller, pwu, hfunk, pwilson}@sift.info

Dr. Lewis Johnson
CARTE Labs USC/ISI
Address: 4676 Admiralty Way
Marina del Rey, CA 90292
(310) 448-8210
johnson@isi.edu

Keywords:
Social Interaction Behavior, Etiquette, Politeness, Avatar,
NPCs, Cross Cultural Training, Training Games

ABSTRACT: *Characters or agents which react appropriately—-taking offense when reasonable, giving deference where appropriate, etc.-- are a fundamental need for believability and accuracy in simulations of social interactions (including culture-specific and multi-cultural interactions). This is especially true for applications where complex and realistic interactions with intelligent agents are important-- such as cross-cultural training for military personnel. We have developed a quantitative, computational implementation of a rich, universal theory of human-human "politeness" behaviors and the culture-specific interpretive frameworks for them (labeled "etiquette") from sociology, linguistics and anthropology. This model links observable and inferred aspects of power and familiarity relationships, the degree of imposition of an act (each of which have implications for roles and intents) and the actor's character to produce expectations about politeness behaviors. By using observations of politeness behaviors (or their lack), the same model permits inferences and updates about those attributes. We present the algorithm we have developed and describe its results in scoring the degree of politeness or rudeness across 8 test cases. We see applicability of this model to interactive agent behavior generation and adaptation through the creation of modular, cross-cultural etiquette libraries. While other methods of interactive behavior generation are available (e.g., behavior scripting) our modular, computational approach should provide substantial payoffs in terms of reducing software development costs and/or increasing the breadth of an agent's social interaction behaviors.*

1. Introduction

The role of social interactions—that is, interactions based on the social characteristics and assumptions of each agent being an intentional entity (Dennett 1989) and drawing from culturally familiar patterns of expectations about appropriate behaviors among such agents—between humans and machines is receiving increasing attention (e.g., Preece, 2002; Miller, 2004) as machine and automation capabilities become more complex and more sophisticated. Similarly, awareness of the importance of culture-specific interaction patterns in multi-cultural human-human interaction (e.g., Hofstede, 2001) is driving an increased need for simulation of culture-specific socially interactive agents for training purposes in both military and commercial applications (Chatham & Braddock, 2003).

Yet accurate models of detailed and extensive social interactions are rare in simulation development efforts. Rarer still are models which can support multi-cultural interactions. And most such models which do exist involve some form of scripted interactions—which have the drawback that they are costly to produce and encode and are generally "brittle"—supporting interactions along only a narrow, pre-defined path with minimal variations.

It has rarely been cost effective to encode simulations which support anything near the flexibility and breadth of true human-human social interactions. Nevertheless, to achieve the goal of interesting and effective training through social interactions in a game or simulation context will require that the agents used in training be “believable” both in the social interactions they exhibit (which must, in turn, be accurate with regards to the culture the agent is intended to be a member of) and in the breadth of actions which an agent of that sort would exhibit or could recognize and respond to.

Computerized Non-Player Characters (NPCs) don’t currently behave with the richness and fluency of social interaction behavior that we expect of them and are therefore, unbelievable in key ways. For example, it is entirely possible, in most “first-person” games which support any form of face-to-face social interaction besides combat, for me to insult non-player characters in a wide variety of ways with no response on their part except in those rare instances where I trigger a script through the use of a key word.

Failures in achieving believable behavior are arguably much more significant for simple “moves” in social interactions—what we refer to broadly as social interaction etiquette—than they are for “unbelievable” appearance or physical movement. We find it quite possible to interact with, trust and even be taught by humans whose appearance and movements are abnormal (perhaps through birth defects or injury). We even rapidly evolve methods for interacting with humans who have no “appearance” at all—for example, when interacting over radio or telephone communication, or in print. But we find it more difficult to interact with machines which fail to behave in accordance with our social rules (Reeves and Nass, 1996) for such matters as who should speak when, what sorts of information should be provided and which should be reserved until requested, who gets to dictate tasks to be performed, etc. And we have seen the rapid evolution of “etiquettes” for interacting with new ”faceless” human-human mediation technologies such as voice mail, email and chat, much of which preserves means of conveying politeness by making it explicit—such as the ubiquitous smiley faces in email communications.

One application for simulated agents that exhibit accurate, believable culture-specific social interaction behaviors is in training soldiers for cultural awareness. It is becoming apparent that such training is important in assisting soldiers to work with local authorities and civilians. The Nobel Peace Laureate Forum has designated Religious and Cultural Conflicts as one of its five major issues. The DoD is well aware of problems that arise from cultural misunderstandings, as well as the value of educating troops in language and social interaction prior to sending them abroad. One specific focus is cross-cultural training (CCT). An NPC which displays social characteristics consistent with its cultural background can provide CCT in an appropriate and cost-effective manner. For example, the Peace Operations Training Center hosted more than 200 soldiers from Fort Hood for a course on Arabic culture in early November of 2003 to prepare them for deployment to Jordan. The current state of cultural training involves foreign instructors covering everything from basic language to dealing with Arabian women during checkpoint inspections (Mares, 2003). While this is an excellent way to introduce the culture, it is resource-intensive, only available to a limited number of soldiers, and provides very little interaction between a trainee and a Jordanian civilian. Given the limitations in human resources required to provide such training, a computer-based NPC may be the only viable solution.

Accurately simulating cultural differences in social interactions requires “socially-aware” agents. Such agents take offense believably if not addressed in a culturally appropriate fashion, might appear recalcitrant or ignorant when they are merely trying to follow their culturally-derived notions of polite turn-taking in discourse, etc. Relevant social interaction behaviors, even those for different cultures and contexts, can frequently be emulated in hand-written scripts and simple, locally-relevant rules. But such approaches are time- and labor-intensive in their own right and brittle--only limited interaction complexity can be supported if every move has to be hand-scripted in advance. A general theory and model of social interactions would greatly enhance the usability and sophistication of NPCs, while improving the speed and/or reducing the cost of their construction.

Therefore, our focus is on developing general models and methods of achieving and assessing believable social interactions between individuals and small groups. We are leveraging existing theoretical work by transferring sound socio-anthropological research on social interactions to develop a computational model to adapt and/or score the interaction behavior of a computer-based NPC in a given role and with a given action intent, as described below.

2. “Politeness” for Social Interactions?

The terms “etiquette” and “politeness” are likely to evoke notions of formal courtesies and which dinner fork to use. But politeness is a technical term and a well-studied phenomenon in anthropology, sociology and linguistics having to do with the processes by which we determine and manage the “threat” inherent in communication and interaction between two intentional agents in a social interaction—that is, agents that are presumed to have goals and the potential to take offense at having those goals thwarted in any interaction where those intentional attributes are relevant (cf. Dennet, 1989; Goffman, 1967). As we see below, politeness in this sense is the method by which we signal, interpret, maintain and alter power relationships, familiarity relationships and interpretations of the degree of imposition or urgency of an act.

We use the term etiquette to refer to the set of expectations about observable behaviors that allow interpretations to be made, in a cultural context, about those who do or do not exhibit them. Observable behaviors are interpreted against a framework of etiquette expectations to allow conclusions about the politeness of those we interact with, while simultaneously, we choose behaviors (consciously or unconsciously) on the basis of the same etiquette framework--which dictates how they will be interpreted by those who observe them. As such, the formal and prescriptive etiquettes formulated by Miss Manners and Emily Post are a particularly stilted type of etiquette, but hardly the only one; more common are the unwritten (and descriptive) etiquettes we encounter, manipulate and react to as we move through our lives—the etiquettes of the classroom, the locker room, the marketplace, etc. Etiquette refers to the expected "moves" in context that allow participants to make inferences about group membership, power relationships, formality/informality, degree of friendship, importance of information conveyed, etc. Violation of etiquette can convey lack of regard, lack of acceptance of the proposed relationships, or can convey overriding concerns such as a critical threat.

Etiquette enables the interpretation of observable behaviors—and thus it makes use of a wide range of verbal, physical, gestural and even more primitive modes of interaction. For example, deference can be expressed by posture, by quiet speech and/or by explicit markers such as titles and honorifics. The key is the set of cultural expectations which allow interactants to interpret the behavior, or lack of behavior, in a predetermined fashion. In this sense, there is a "cultural etiquette" associated with, say, infantry soldiers as opposed to clerical workers, just as there is a one for marketplace negotiations in the Middle East vs. an American shopping center.

As such, therefore, politeness and etiquette are very much at the forefront of determining the believability and effectiveness of NPCs engaged in interactions with other social actors in training applications in militarily relevant domains. Believable behavior is behavior that is understandable (i.e., the viewer can infer intent behind the behavior) and broadly consistent with the viewer's expectations. Understandability and expectations, in turn, depend upon the social and cultural context of the behavior. Etiquette provides a way of modeling interactions and moves within a social and cultural context, and of predicting their impact on observers' interpretations about the motives, understanding, knowledge and relationships of those who exhibit them. As we will develop below, believability in social interactions means behaving in accordance with expectations for an actor who knows the social conventions and has a personal stake (personal goals to be thwarted) in the outcomes. Therefore, we focus in this project on etiquette and its role in achieving believability. If NPCs do not behave in accordance with etiquette-based expectations, one of two outcomes may result: either (1) they will not be perceived as believable, or (2) they will be misinterpreted—the trainee will draw false inferences about their relationships, intentions, etc. In either event, they will be useless for training purposes—and worse yet, they may produce inaccurate expectations in students who interact with them.

3. A Model of Human-Human Etiquette for Politeness

A seminal body of work in the sociological and linguistic study of politeness is the cross-cultural studies and resulting model developed by Brown and Levinson (1987). Brown and Levinson noted that people across cultures and languages very regularly depart from strictly efficient conversation by using an array of conversational behaviors designed to mitigate or soften direct expressions of desire, intent or command. A simple example in English will illustrate the point: as we settle down to a meal together and I ask you "Please pass the salt," the use of "please" in that sentence is unnecessary for a truthful, relevant or clear expression of my wish and is, in fact, an explicit addition of verbiage not required to express my intent (to have the salt passed to me).

Over years of cross linguistic and cross cultural studies, Brown and Levinson collected and catalogued a huge database of such violations of efficient conversation. Their explanation for many of these violations is embodied in their model of politeness, which will be explained next.

3.1 Face threats in social interactions

The Brown and Levinson model assumes that social actors are motivated by two important social wants based on the concept of face (Goffman, 1967) or, loosely, the "positive social value a person effectively claims for himself" (cf. Cassell and Bickmore, 2002, p. 6). Face can be "saved" or lost, and it can be threatened or conserved in interactions. Brown and Levinson further refine the concept of face into two specific subgoals that all social actors can be presumed to have:

1. *Positive face*—an individual's desire to be held in high esteem, to have his/her actions and opinions valued, to be approved of by others, etc.
2. *Negative face*—an individual's desire for autonomy, to have his/her will, to direct his/her attention where and when desired, etc.

Virtually all interactions between social agents involve some degree of what Brown and Levinson call Face Threatening Acts (FTAs). My simple act of speaking to you, regardless of the content of my words, places a demand on your attention that threatens your negative face, for example. This, then, is the reason for the "please" in my request for salt: If I simply state my desire that you give me the salt as bald propositional content (e.g., "Give me the salt") I may efficiently communicate that intent, but I have also been ambiguous about whether or not I have the power or right or can otherwise compel you to give me salt. You might well take offense at the implication that I could demand salt from you.

The "please" in the example above is an example of a politeness strategy used to "redress" or mitigate the threat contained in the request for the salt. Furthermore, the expectation that such a strategy be used in certain contexts is an example of etiquette that enables interpretations. The etiquette which we believe to be in play entitles us to conclude that those who use "please" in an appropriate context are striving to play by the rules—striving to be seen as polite; those who do not are not striving to be polite for various reasons (perhaps they don't believe they need to be, perhaps their notions about politeness are different, perhaps they are just rude).

3.2 Computing the severity of a face threat

The core of Brown and Levinson's model is the claim that the degree of face threat posed by an act is provided by the function:

$$W_x = D(S,H) + P(H,S) + R_x$$

- ❑ Wx is the 'weightiness' or severity of the FTA x

- ❑ D(S,H) is the social distance between the speaker (S) and the hearer (H). Social distance is roughly equivalent to familiarity—it increases with contact and interaction, but may also with be based on a priori factors such as membership in the same family, clan or organization and perhaps on being in a "familiar" setting as opposed to a formal one—a sporting event rather than a church. Social distance is a symmetrical relationship—S and H share the same social distance. In training contexts, social distance might derive from familial or clan relationships among characters, or it might be used to convey (or invite) a deeper degree of familiarity with an NPC tutor, sidekick or counselor.

- ❑ P(H,S) is the relative power that H has over S. Power comes from different sources in different cultures and organizations. Clearly, a tutor needs to maintain some power over a student, but NPCs representing commanders, subordinates, or high or low status citizens might all need to act, and to be handled according to different etiquettes if face threats are to be minimized. Power is an asymmetric relationship between S and H.

- ❑ R_x is the ranked imposition of the raw act itself. Some degree of imposition is culturally defined—it may be inherently more of an imposition to request food from a host in Western culture than in an Arabic one, for example. But imposition is also dependent upon the roles and duties of the parties involved. One reason a tutor can correct a pupil, even though s/he might have lower power in the society, is that the correction is expected from the tutor and is, therefore, less of an imposition.

Brown and Levinson themselves do not operationalize these parameters; instead, they are offered as qualitative constructs. Recent work by Cassell and Bickmore (2002) and by Johnson (2004) has created numerical representations for them. In Cassell and Bickmore's work, the resulting computational model was used as a component in a conversational agent (a real estate salesperson) whose goal is to use small talk to increase familiarity to the point where a more face threatening conversational topic (such as personal income level) can be introduced. Johnson has used a similar model to create a pedagogical agent that is designed to maintain and enhance learner confidence and motivation, by offering advice and criticism in ways that protect the learner's face (cf. Johnson, 2004; Wang, et al., 2005). Our goal has been to develop a computational formulation of the Brown and Levinson algorithm for use in free-flowing conversation and social interactions between humans and agents in a simulation environment.

3.3 Redressing face threats

Since FTAs are potentially disruptive to human-human relationships, we generally make use of redressive strategies to mitigate the degree of face threat imposed by our actions. Brown and Levinson offer an extensive catalogue of universal strategies for redressing, organized according to 5 broad strategies. These are illustrated in Figure 1 ranked from least to most threatening.

- ❑ The least threatening approach is simply not to do the FTA. At some threshold, in some contexts and cultures, it will simply be too threatening for some FTAs to be performed, regardless of the amount of redress offered. At this point, the only viable strategy is to avoid doing the act.

- ❑ If one is to do the FTA at all, then the least threatening way to do it is “off record”. Off record FTA strategies are means of doing the act with a sort of “plausible deniability” by means of innuendo and hints. An “off record” method of asking for salt might be “I find this food a bit bland.”

- ❑ If one does FTA overtly, then one can still undercut its degree of threat by offering redress aimed at either positive or negative face. Brown and Levinson suggest that negative redress will be more effective (less threatening) than positive. Negative redressive strategies focus on H’s negative face needs—independence of action and attention. They minimize the impact on H by being direct and simple in making the request, offering apologies and deference, minimizing the magnitude of the imposition and/or explicitly incurring a debt. “I’m sorry, but I’d be very grateful if you could just pass me the salt” includes many negative redress strategies (apology, incurred debt, minimization of the imposition).

- ❑ Positive redressive strategies target the hearer’s positive face needs—the desire that his/her needs and wants be seen as desirable. These strategies emphasize common ground between S and H by noticing and attending to H, by invoking in-group identity, by joking and assuming agreement and/or by explicitly offering rewards/promises. “Hey buddy, you want to pass me that salt, don’t you?” uses positive redressive strategies including both an in-group identity marker and assumed compliance.

- ❑ Finally, the most threatening way to perform an FTA is “baldly, on record,” without any form of redress. In some cases where power of S over H is high, familiarity is high and/or imposition is low, doing an FTA with no redress may be the expected thing to do. The “Give me the salt” example used above is a bald, unredressed form of performing that FTA.

Brown and Levinson’s model doesn’t stop at that level, however. For positive and negative redressive and off record strategies, they offer a host of well-researched examples from at least three different language/culture groups (English, Tamil and Tzeltal) organized into a structure of mutually supporting and incompatible approaches. We do not have space to present their findings in depth, but we note as an example that their categorization of negative redress strategies contains 10 alternate approaches, some of which are mutually supporting or conflicting, including:

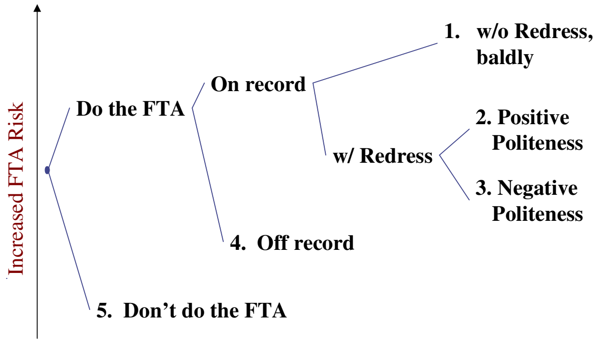

Figure 1. Universal redress strategies as ranked by Brown and Levinson (1987).

- ❑ *Be Pessimistic*—“You’re not going to pass me the salt, are you?”
- ❑ *Minimize the Imposition*—“Could you just nudge that salt shaker over here?”
- ❑ *Give Deference*—“Excuse me, sir, would you pass the salt?”
- ❑ *Apologize*—“I’m sorry to interrupt, but would you pass the salt?”

4. An “Etiquette Quotient”—Believable levels of Politeness

According to the Brown and Levinson model described above, people generally want to accomplish their goals expeditiously-- and this argues for minimizing redressive strategies. But they also experience a range of social and personal pressures to not threaten the face of those they interact with (especially those with greater power or shared familiarity)-- and this argues for extensive redressive strategies. The balance between these pressures yields the selection of specific strategies in context. Brown and Levinson allude to, but don’t explicitly include a factor representing the relative weighting that an individual puts on his/her own goals vs. the face goals of others-- his/her general willingness (independent of the other factors) to place others’ needs first. For want of a better term, we'll call that “character” and introduce a term for it, abbreviated as C, with the character of speaker (S) being C(S). In other words, the degree of redress that a speaker S chooses to use will be a function of the degree of face threat inherent in the act (itself a function of P,D and R) and the speaker’s character C(S).

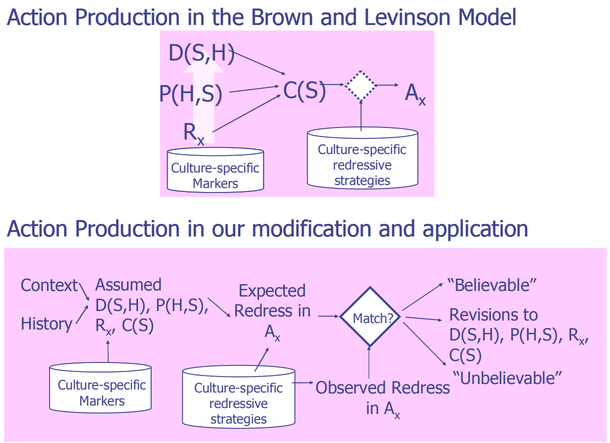
Action Production in the Brown and Levinson Model
D(S,H)
P(H,S)
C(S)
A$_x$
R$_x$
Culture-specific Markers
Culture-specific redressive strategies
Action Production in our modification and application
Context
History
Assumed D(S,H), P(H,S), R$_x$, C(S)
Expected Redress in A$_x$
Match?
"Believable"
Revisions to D(S,H), P(H,S), R$_x$, C(S)
"Unbelievable"
Observed Redress in A$_x$
Culture-specific Markers
Culture-specific redressive strategies

Figure 2. Graphical depiction of the Brown and Levinson model and our proposed modifications to it.

But the above description, and indeed Brown and Levinson's primary focus, is from the perspective of the speaker/actor (S) interested in achieving interaction goals and, presumably, in avoiding face threat to hearers (Hs). We can characterize Brown and Levinson's model graphically as in the top portion of Figure 2. A Speaker with a given character C(S), uses his/her knowledge of the D, P, and R of a given context and desired FTA in order to select one or more strategies from among a knowledge base of culture-specific strategies resulting in a specific action A_x which is designed to both further S's goals and to avoid undue face threat to his/her interlocutors.

In order to implement and make use of this model in believable human-computer interactions (i.e., with simulated agents), we need to take the perspective of an observer/hearer (who may or may not be the one the speaker is actually interacting with). This perspective is represented graphically at the bottom of Figure 2. Here, an observer (O) perceives an utterance that has bald content as a speech act and may or may not contain culturally-recognized redressive strategies. O also has access to additional cues from his/her perception of the context and perhaps memory for past history. Given these cues, O's goal is to construct a picture of the "politeness" character of S and, through that, to the P, D and R of the interaction between S & H. Fundamental to our approach is the claim that this construction process is based largely on the degree of match or mismatch between the redressive strategies actually used by S (as perceived by O) and those expected by O.

Given his/her own observations or knowledge of the context, O can construct an understanding of the parameters P, D, and R. For example, if S is noticeably older, richer, or is wearing insignia that make it clear that s/he outranks H, then O might reasonably conclude that the power distance (P) between them is large and favors S. If S and H are behaving familiarly (standing close together, interacting jovially, using nicknames, etc.), are known to be related as family members or friends, etc., then O might conclude that the social distance (D) is comparatively small between them. Finally, O will have his/her own culturally-based beliefs about the degree of imposition (R) of a given act (e.g., asking for money is a greater imposition than asking for help finding a location, which is a greater imposition than asking for the time), but observed or known characteristics of the interaction may also serve to reduce the perceived R. For example, if S is known to have a duty (perhaps based on his/her role) or a standing request to provide certain information or advice to H, or if H is not apparently engaged in any ongoing activity.

Then, given his/her beliefs about these parameters, O can construct an estimate of the degree of face threat associated with the bald content of the act. Furthermore, given whatever information s/he possess about C(S), O can adjust his/her predictions about the degree, and therefore the types, of redressive actions that s/he might expect to see used. Let us call this product the *expected act* (A_x).

But at the same time, O can actually perceive an *observed act* (A_x). S performs an act that O will perceive as having a degree of imposition and, perhaps, various associated redressive actions. If the observed act and the expected act are the same (perhaps within certain degrees of tolerance), then the actor will be seen as believable—at least with regards to his/her/its politeness-producing etiquette behaviors.

Therefore, one metric for believability is the delta between the expected act and the perceived act. And yet, other humans fail to behave as we expect them to behave all the time without our labeling them "unbelievable". This seems to be because humans are generally aware that predicting politeness behaviors is far from an exact science. We are generally more willing to revise our beliefs about aspects of the context or character that produced our initial predictions and then reassess that prediction than we are to conclude that S is acting artificially. This metric may be computed over time as well. If successive actions, with their associated degrees of redress employed, continue to violate O's notions of the NPCs' context and P,D,R and C values, O may choose to revise the assumed characteristics seeking a set of P,D,R and C values that minimizes the delta between expected and observed degrees of redress. If no such model is found, or if violations are extreme, s/he may give the game up and simply declare those behaviors to be "unbelievable".

5. Algorithm Implementation and Test Cases

In work funded by a DARPA Small Business Innovation Research grant, we have recently completed the initial development of an "Etiquette Quotient" (EQ) algorithm based on Brown and Levinson's work as described above and have demonstrated its capability to provide expected politeness assessments of eight test cases. Our approach and results will be described in this section.

5.1 The EQ Algorithm

We have implemented a version of the Brown and Levinson equation to use as a predictive model of the believability of the redressive actions of a computerized game character (a Non Player Character or NPC) as it appears to a human observer, with perceived aspects of context (D, P and R, as well as known history about the character, C, of the actors). NPCs which do not exhibit the expected degree of polite redress (either by being over- or under-polite) are expected to be seen as either unbelievable or to invite rethinking of what was previously understood about the D,P,R and C of the context. For example, if a private bursts in on his captain and issues a bald directive ("Get your coat on") without any redress, an observer might well assume that the degree of imposition (R) is less than might otherwise be the case because the private was charged with giving such instructions, or that the familiarity between them warranted it. Otherwise (and especially in a simulated environment), the observer might simply believe that the private was behaving "unbelievably".

Expanding on the Brown and Levinson equation, our implementation uses weights on each component to allow the possibility to value D, P, and R differently (a factor we suspect may underlie some cultural differences), resulting in the equation below:

$$W_x = [w_1 \cdot D(S,H)] + [w_2 \cdot P(H,S)] + [w_3 \cdot Rx] + C(S)$$

Each Observer adds his/her own interpretations of the context. For example, D(S,H) could be expanded to $[B_H{:}w_1 \cdot B_H{:}D(S,H)]$, representing Hearer's belief about the degree of social distance and the Hearer's belief about the appropriate weight for the social distance term. We use Speaker belief (B_S:) and Observer (who could also be a Speaker or Hearer) belief (B_O:) similarly. This results in the following expansion for an Observer O:

$$B_O{:}W_x = \{[B_O{:}w_1 \cdot B_O{:}D(S,H)] + [B_O{:}w_2 \cdot B_O{:}P(H,S)] + [B_O{:}w_3 \cdot B_O{:}R_x]\} + B_O{:}C(S)$$

An implication of the Brown and Levinson model, though never overtly spelled out, is that the weightiness of the face threat must be fully compensated for, or "redressed" in normal interactions if the status quo in relationships is to be maintained. Therefore, the value of W_x should be balanced by the "value" (V) of a set of redressive actions used in the interaction x ($\mathbf{A_x}$) if the interaction is to appear "normal" or believable or without ulterior motive. In other words, we expect the value of the redressive strategies the speaker uses to equal or balance the value of the face threat s/he produces, or:

$$W_x = V(\mathbf{A_x})$$

This means that an Observer's beliefs and weightings of the social distance, power and imposition relationships, adjusted by belief about the character of the Speaker, should be balanced by the Observer's belief about the value of the set of redressive behaviors used. We express this as a difference to give us an "incredibility" or "imbalance" metric which also serves as a perceived politeness metric:

$$B_O{:}I_x = \quad B_O{:}V(\mathbf{A_x}) - B_O{:}W_x$$

In order to use this metric to evaluate the imbalance between expected and observed levels of politeness, we must operationalize the various parameters. Space does not permit a detailed presentation of our method for accomplishing this, but we will summarize the basic approach below.

5.2 Operationalizing EQ Terms

In order to operationalize and quantify the Brown and Levinson model described above, we first developed scalar values for the various politeness parameters P,D, R and C. We proposed that the variables D(S,H), P(H,S) and Rx, as well as the various parties' perceptions of them, be represented as continuous scalar values ranging from negative to positive infinity. The value of 0 is the "balance point" or a nominal or equal value for each scale, while positive values indicate that the dimension is increased (and contributes to an increasingly "weighty" or potent FTA) and negative values indicate that it is decreased (and is, in fact, building up the Hearer rather than threatening him or her). For Power Difference of the Hearer over the Speaker (P(H,S)), for example, a value of 0 means that the power of the Hearer and the Speaker are equal, that they are (exact) peers. Values greater than 0 indicate that the Hearer (H) has increasingly greater power (as values increase) relative to S and, therefore, that face threat increases whenever S addresses H. We proposed the following scale anchor points for nominal American culture:

- A value of -1000 is characteristic of the power that a CEO of a major company (as S) has "over" (or relative to) a janitor in his/her company (as H) or the power that a parent has over a small child.
- A value of -100 is characteristic of the power that a professor has relative to a freshman student or a parent over an early teenage child.
- A value of -10 is characteristic of the power that a project manager in an informal research team has over project members, or the power that a parent has over an older teenager.
- A value of 0 is characteristic of equal power between S and H; no or negligible difference—for example, the power relationship between two co-workers at the same level and seniority.
- A value of 10 is characteristic of the inverse of the power described for -10 above—the power that an older teenager or work team project member as S would have over (or relative to) a parent or project manage, as H, respectively.
- Etc.

Similar scales were developed for D(S,H) and Rx. The character term (C) was represented as a simple value added or subtracted from the W_x sum.

Next, we developed numerical valuations for various redressive behaviors based on the guidelines provided by Brown and Levinson as depicted in Figure 1 (illustrating the various redressive values of broad classes of redressive strategies) above. Ranges of values for the broad classes of strategies were defined as follows, with individual strategies within each class being assigned a value within the designated range:

- The value of the use of an individual positive redressive strategy (see Brown and Levinson, p. 102, Figure 3, for a list of such strategies) will provide from 1 to 40 "units" of redress.
- The value of the use of an individual negative redressive strategy (see Brown and Levinson, p. 131, Figure 4, for a list of such strategies) will provide from 20 to 60 units of redress.

Within the range defined above, a specific score was assigned to individual instances of redress which fell into the category, as will be illustrated below.

The effects of multiple redressive strategies were scored as simply additive. We understand that this is a simplification, and that the efficacy of added redressive behaviors probably falls off, eventually becoming simply irritating. This means that the value V of a set of N redressive actions A contained in interaction x is given by the function:

$$V(\mathbf{A_x}) = V(\mathbf{A_1}) + V(\mathbf{A_2}) + \ldots\ V(\mathbf{A_N})$$

5.3 Evaluation Test Cases

This approach was then tested in a series of sample social interaction vignettes crafted to represent (according to our American cultural intuitions) either normal/balanced politeness, unbelievable over-politeness or unbelievable rudeness. Our goal was to determine if the equation and scoring techniques we had developed would track our intuitions for scenarios in which politness was balanced or imbalanced in various ways. The level of face threat and redress were varied over this set of vignettes so that high face threat situations were paired with high levels of redress (which should produce roughly normal, balanced levels of redress) as well as low levels of redress (which should be highly imbalanced and perceived as very rude—perhaps unbelievably so). Similarly, very low levels of face threat were paired with very high levels of redress (which should be perceived as over-polite, perhaps unbelievably so) and with low levels of redress (which should be perceived as balanced and fully believable). Examples of two such vignettes are illustrated below:

Vignette 1 —*High Face Threat, High Redress, High Believability*: A low ranking soldier (i.e., a corporal, as indicated by uniform insignia) walks into the Mayor's office and the Mayor motions him quickly to a seat. The soldier takes off his hat and sits down, waiting while the Mayor continues to write something. The Mayor finishes up writing, puts down his pen and looks up at the soldier expectantly. The soldier then says, "I'm sorry to interrupt you work, Mayor Fredrickson, but my name is Corporal Jones and I've been put in charge of your escort to the event tonight. I was wondering if it would be possible for you to let me know where I can meet your wife so that I can get her there on time?"

Vignette 2—*High Face Threat, Low Redress, Low Believability*: As for vignette 1 above except that the soldier acts and speaks differently. Here, he interrupts the mayor while he is speaking, perhaps by putting a hand on his shoulder, and says loudly, "Tell me where I can meet your wife?"

Each of the eight vignettes was then assessed using the operational scoring tables we had created. For example, for the first vignette the imbalance evaluation proceeded as follows:

- The corporal (as S) has lower power than the mayor by a fairly large degree. That is, his "power difference" relative to the mayor is fairly large—perhaps slightly larger than our anchor point of 100, yet less than the anchor point of 1000. We scored

Table 1. Redressive actions scored in Vignette 1.

Action and Interpretation	Score
1. The soldier waits until the mayor is finished and invites him to speak. This seems to be a very explicit form of negative politeness (putting the other's interests first) and, especially in this instance where the H was not actively engaged in another conversation, seems very potent.	60
2. The soldier takes off his hat. This is a sign of deference in our culture, which is in turn a fairly potent negative politeness strategy.	50
3. The soldier apologizes for interrupting. This is also a negative politeness strategy, though arguably a less potent one (though that may be highly mitigated by facial expressions and body language).	30
4. The soldier uses an honorific. Moderately potent negative politeness strategy.	40
5. The soldier poses the FTA as a question. Common negative politeness strategy.	20
6. The soldier offers an explanation/reason for needing the information. Positive politeness strategy, seems particularly powerful in this case.	35
7. The soldier appeals to the Mayor's (H's) interests. Positive politeness strategy Powerful in this context.	30
8. The soldier is hesitant and skeptical about compliance. A common but reasonably potent negative politeness strategy.	30
TOTAL	295

that as P(H,S) = 300 (and, since there were no cultural differences or speaker or observer misperceptions B_O:P(H,S) = 300).

- There is no particular familiarity between the two individuals in this vignette, but social distance is not extreme either. They are from slightly different "cultures" (military vs. civilian infrastructures) and show no evidence of prior relationship, but they are engaged in a common endeavor. The social distance between them is probably only slightly higher than 0. Thus, D(S,H) = 3.
- The imposition of this request could be somewhat large. To ask after the location of one's wife so as to pick her up is comparatively threatening, though the fact that this is in the mayor's service should mitigate this imposition (as the corporal reminds him). The raw imposition is a short answer required from the mayor, characteristic of our level 10, so we assigned it: $R_x = 10$.
- Since we have provided no reason to believe that the character of the corporal is anything other than nominal, we will assume that C=0.

This gives us a value of $B_O:W_x$ as supplied by the left hand portion of the equation above as:

$$3 + 300 + 10 = 313$$

For the value of the redress applied $V(A_x)$ we identified and scored the set of redressive actions in Table 1.

Thus, the imbalance score for this vignette, as calculated by equation 6, would be: 295 – 313 = -18. Since this vignette was intended to convey both high face threat and high redress and, thus, to be balanced and believable, this score seems to be about right, falling very near zero.

For the second vignette, by contrast, we have a high degree of face threat with virtually no redressive actions. This is unexpected and should be perceived as very rude. This scenario should have a score much less than 0 on our imbalance metric—indicating that there is substantial unredressed threat. This vignette was scored as follows:

For the degree of face threat (Wx), we scored the following:

- The power relationship is identical to the above P(S,H) = 300.
- The social distance between them should also be identical to the above.
- The imposition of this request could be large, especially in the absence of the reminder that it is in the mayor's interest, but it remains essentially the same request as in the previous vignette. Thus, we keep the imposition score the same: Rx = 10.

This gives us the same Wx score as in the first vignette above: 3 + 300 + 10 = 313.

Table 2. Redressive actions scored in Vignette 2.

Action and Interpretation	Score
1. The soldier is very brief and therefore, takes little of the mayor's time. This could be counted as a negative politeness strategy of directness (albeit not a very effective or unambiguous one).	30
2. This could perhaps also be counted as an example of the positive politeness strategy of optimism and assumed compliance—though again, not unambiguously.	10
TOTAL	40

For the value of the redress applied V($\mathbf{A}_x$) we see none of the redressive strategies used above. At best, the redressive strategies illustrated in Table 2 are used.

Thus, the imbalance score for this vignette, from equation 6, would be: 40 - 313 = -273. This again seems to be about right—a score much less than zero for a vignette that was intended to include much more threat than redress.

An evaluation similar to that described above was carried out for a total of eight vignettes and the quantitative algorithm tracked predictions for rude, polite or nominal perceived etiquette levels very closely. As shown in Figure 3, all vignettes that were intended as "nominal" interactions (that is, using about the amount of redress as would be expected in American culture for the amount of redress offered) scored within +/- 100 points of zero. All vignettes that were expected to be seen as over-polite scored well higher than 100 points; while all that were expected to be seen as overly rude scored substantially less than -100 points.

While the above example was based on one individual's scoring assessments (Dr. Miller's), we have since replicated this work with two other "raters" following a brief training session from Dr. Miller. Each rater then scored the vignettes according to the guidelines as described above (and in more detail in our training documentation). The results of the three raters' scores for the eight etiquette scenarios were then statistically compared. The top-level imbalance metric (Ix) showed a Robinson's A correlation of .931 among the three raters across the 8 vignettes, and the two major subfactors (Face Threat Weight—W_x and composite Redress Value—V($\mathbf{A}_x$)) the Robinson's A correlations were.950 and .863 respectively. These values are all well above traditional correlation thresholds of .7 or .8 for multiple judge rating correlations. Thus, this study lends weight to the belief that we have identified a reliable method of scoring the degree of politeness vs. expectations in social discourse—at least within an American cultural setting.

6. Conclusions and Future Work

An ability to score the believability of the social interaction behaviors of an NPC is important because it allows for machine reasoning about how social interaction "moves" will be perceived. Thus, when fully operationalized, this algorithm holds the potential to become a general, reusable, computational approach to equipping an NPC with the ability to evaluate the human player's actions given what it knows about P,D,R and C and therefore, to reactively take offense or take advantage. Similarly, it can equip NPCs with the ability to determine appropriate behaviors to exhibit in order to further their ends.

The use of Brown and Levinson's model and theory to inform a module for reasoning about social interaction behaviors guarantees that the module will be universal in its reasoning about and scoring of abstract politeness "moves". Any such module will need to be equipped with culture-specific knowledge bases, however, to enable reasoning from the observable behaviors in a culture (e.g., pursed lips or a rigid hand-to-eyebrow salute) to the abstract etiquette "moves" (and therefore, politeness implications) over which the model's parameters are scored. This has the practical implication that the general social interaction reasoning of an NPC or other simulation can be effectively modularized, and, thus, large savings in simulation code development can be realized. Furthermore, basic game storylines or training modules and even specific characters can be easily transposed from one cultural milieu to another—enabling the village priest who the player had to interact with to get intelligence information in a Kosovo training game to take on the culture-specific behaviors and reactions of an imam in the Iraq training game. In each case, new knowledge bases of culture-specific politeness behaviors would need to be developed (and, of course, checked for accuracy) for each new game, but the core game storyline(s) and character roles, general actions, motivations, capabilities, etc., could remain unchanged.

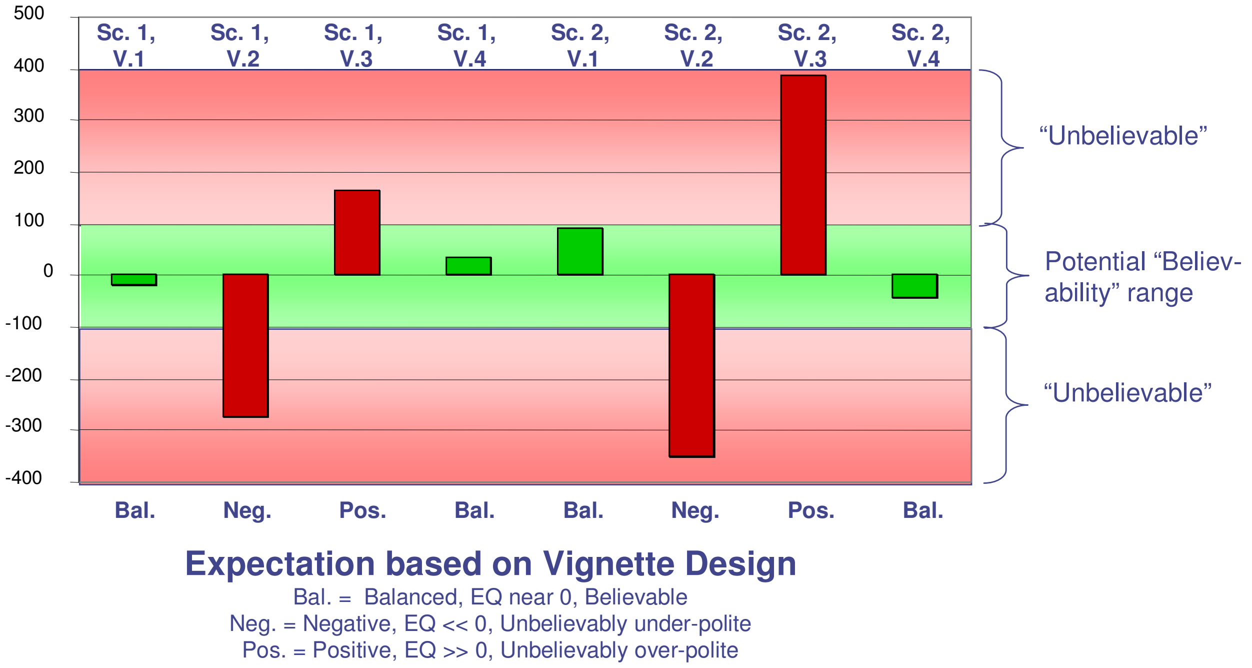

Figure 3. Imbalance scores for all eight vignettes.

We are currently engaged in a follow on experiment to further tune and validate the predictions of the algorithm by exploring the set of vignettes described above with a much larger audience of American college students. Concurrently, we are working to integrate our algorithm and its associated representation and reasoning approach into a social interaction simulation involving multiple NPCs and to enable the NPCs to ascertain the relative threat and redress of behaviors directed at them, as well as generating redressive behaviors in keeping with their goals and relationships with a human player or trainee. Beyond that, we hope to begin work on developing "culture modules"—the representation and culture-specific knowledge to be integrated into our basic Brown and Levinson-based computational algorithm to give the resulting NPCs the specific library of observable cultural redressive behaviors and sensitivities to behaviors directed at them that they might possess if they were, say, Iraqi, Kosovar, German or French.

The outcome of our proposed developments will be a dramatic increase in the ability to rapidly create computer training simulations or games with realistic, culture-specific social interaction models for their NPCs. We anticipate at least a 10x reduction in the time required to generate equivalently rich social interactions included in a 30 minute game episode as compared to the use of traditional scripting approaches. Our approach will also provide for much greater flexibility in the interactions which NPCs can exhibit—and thus, richer interaction capabilities which will, by allowing greater and more user-driven exploration capabilities, have payoffs in terms of the engagement and training effects in the applications in which they are used.

7. Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency and U.S. Army Aviation and Missile Command under contract number W31P4Q-04-C-R221. We would like to thank Dr. Ralph Chatham, our Program Monitor.

This document was "approved for public release, distribution unlimited" by the DARPA Technical Information Office, on March 20, 2006.

8. REFERENCES

Brown, P. & Levinson, S. (1987). *Politeness: Some Universals in Language Usage*. Cambridge,UK.; Cambridge Univ. Press.

Cassell, J. and T. Bickmore. (2003). Negotiated Collusion: Modeling Social Language and its Relationship Effects in Intelligent Agents. *User Modeling and User-Adapted Interaction. 13(1):* 89-132

Chatham, R., and J. Braddock, (2003). *Training for Future Conflicts*, Report of the Defense Science Board.

Dennet, D., (1989). *The Intentional Stance*, Cambridge, MA; MIT Press.

Goffman, E. (1967). *Interaction Ritual: Essays on Face to Face Behavior.* Garden City; New York.

Hofstede, G. (2001). *Cultures and Consequences, 2nd Ed*. Thousand Oaks, CA; Sage Publications.

Johnson, L., Rizzo, P. (2004). Politeness in Tutoring Dialogs: "Run the Factory, That's What I'd Do". *Intelligent Tutoring Systems,* 67-76.

Mares, Pvt. K. L. (2003). Jordanian course preps soldiers on Arabic culture, *Army News Service*, Nov. 10, 2003. Retrieved from the web April 19, 2006: http://www4.army.mil/ocpa/read.php?story_id_key=5395

Miller, C. A. (Ed.), (2004). Human-Computer Etiquette. *Communications of the ACM, 47(4)*. 30-61.

Reeves, B. and Nass, C. (1996). *The Media Equation*. Cambridge, UK; Cambridge University Press.

Preece, J. (Ed.) (2002). Supporting community and building social capital. *Communications of the ACM 45(4)*, 36-73.

Wang, N., Johnson, L., Mayor, R., Rizzo, P, Shaw, E., and Collins, H. (2005). The Politeness Effect; Pedagogical Agents and Learning Gains. In *Proceedings of the 12th International Conference on Artificial Intelligence in Education*.