Agent-based modelling to visualise trustworthiness: a socio-technical framework

Shuyuan Mary Ho*

College of Communication and Information,
School of Library and Information Studies,
Florida State University,
142 Collegiate Loop,
Tallahassee, FL 32306-2100, USA
E-mail: smho@fsu.edu
*Corresponding author

Ram Reddy Katukoori

College of Arts and Sciences,
Department of Computer Science,
Florida State University,
142 Collegiate Loop,
Tallahassee, FL 32306-2100, USA
E-mail: rk12@my.fsu.edu

Abstract: This paper describes a socio-technical study based on physical world scenarios of deceptive behaviour occurring in a virtual collaborative environment. An agent-based modelling (ABM) approach was adopted to visualise trustworthiness that can signal deceptive behaviour in virtual communications among social actors. The modelling strategies were guided by attribution theories toward an agent's perceived trustworthiness. The assessment of an agent's trustworthiness is based on their language actions as observable information-based behavioural cues, derived from objective semantic analysis. The consistency between an agent's words and actions and distinctiveness of the agent's behaviour when compared against his/her regular behaviour can serve as input data in each interaction. A set of ABM rules is proposed to systematically capture this interaction and assessment of trustworthiness. The logic behind a dyadic attribution can support computations on the trustworthiness of an agent.

Keywords: trustworthiness; attribution theory; agent-based modelling; ABM; online game simulation; human computer interactions.

Reference to this paper should be made as follows: Ho, S.M. and Katukoori, R.R. (2013) 'Agent-based modelling to visualise trustworthiness: a socio-technical framework', *Int. J. Mobile Network Design and Innovation*, Vol. 5, No. 1, pp.17–27.

Biographical notes: Shuyuan Mary Ho is an Assistant Professor at the School of Library and Information Studies, College of Communication and Information, Florida State University. She received her PhD from the School of Information Studies at Syracuse University, her MBA from University of Hartford, and BS Computer Science from Ohio Dominican University with over 16 years of industry and research experience in information systems security. Her current research is situated within socio-technical studies of trustworthiness attribution, human computer interactions, and information systems security.

Ram Reddy Katukoori is a graduate student at the Department of Computer Science, Florida State University. He is a Research Assistant working with Dr. Shuyuan Mary Ho at the iSensor Laboratory, Florida State University.

1 Introduction

Detecting human factors for trusted communications based on behavioural cues in language actions or psychological profiles is a complex socio-technical problem because it involves recognising potentially deceptive communicative intent that is embedded in the complex web of human computer interactions when typical cues of physical interaction are absent. In this research, we attempt to discover simple set of rules and logic to represent human trustworthiness as observed by human 'sensors' in an environment of computer-mediated technologies (Ho, 2009b).

To illustrate how human 'sensors' attribute behaviour, let us first consider the example of a financial analyst who works for a large multinational organisation. This financial analyst is recognised for his advanced understanding of the bid approval process, and adhering to company protocols. Being a part of this team, his peers are comfortable with him and readily share their thoughts with each other. This analyst is also a highly trusted member of his team. He always engages others on the team to share new ideas, and corrects them wherever they go wrong. It has never been the case that the analyst approved an inappropriately priced bid without consultation amongst the team and higher management spread across the globe. This observation of this analyst's integrity is consistent across different scenarios and under similar circumstances over time. During group discussions on a bid review, the commitment by the analyst (i.e., his words) when deciding on a final quote always translates into corresponding actions of submitting the proposal to the sales team for presentation to the client. This scenario often presents a subtle dilemma for the analyst. The dilemma is situated between the fine lines of the actor's loyalty to his group vs. his own personal gains and interests. The analysts' integrity is attributed to be high and the analysts' team tends to have a general consensus on his trustworthiness, since his actions always conform to his words based on historical behaviours. Since the analyst's actions have always been consistent, it is relatively easy for the group to have consensus toward his observed behaviour inferring the trustworthiness of this individual.

In the real world, people understand and learn about each other primarily through facial expressions, eye contact and body gestures as they converse in physical space. However, they interact differently in virtual environments. A diverse virtual environment is provided by clustered computer-mediated technologies that allow people to interconnect, interact and communicate across geographies, time, space and context. For example, a well-connected virtual team may chat and share their thoughts through Yahoo messenger or Facebook. But patterns of virtual interaction have never been homogeneous, and so it is difficult to capture the dynamics of such stochastic interactions among social actors. When unanticipated scenarios arise, such interaction among the social actors becomes complex (Bonabeau, 2002). In order to understand and assess the impact of a social actor's interactions in unexpected situations, it can be helpful to use an agent-based modelling (aka: ABM) that captures cognitive aspects of how social actors (i.e., agents) interact, act and react in various situations, and how they perceive each other's interactions. Thus, we intend to seek answers to this following question.

> "How do we model social actors' attribution in a small group's virtual interaction to cognitively visualize a targeted agent's trustworthiness, which might have been compromised?"

This paper is outlined as follows. In Section 2, we describe the basic principles of social interaction in a virtual communication environment using the example of a financial analyst. In Section 3, we discuss how trustworthiness is defined. In Section 4, we explore how attribution mechanisms work based on the Kelley et al.'s (1973) ANOVA model. Extending Kelley's attribution theory, we then discuss how dyadic attribution model works in Section 5. We review how ABM techniques can be utilised to model dyadic attribution model in particular for small groups in Section 6. In this section, we also demonstrate how simple rules can be constructed for a dyadic attribution mechanism, which can then be used to mimic how each social actor processes information fed into an agent's team's interactions. However, it is important to note that there are limitations to deploy dyadic attribution model as the systematic analysis engine using ABM approach. The limitation, contribution, and future work of this study are concluded in Section 7.

2 Basic principles in virtual interactions

Interpersonal communication can be considered to be part of an information network involving social entities that are interconnected through relationships (Urban and Schmidt, 2001). Every social actor is an autonomous decision making entity, which is the definition of an agent. Each social actor individually assesses situational context. Based on sets of predefined rules and conditions, he or she makes decisions. We say a social actor A acts and sends messages to another actor who functions as an observer B_i in this information network. The observer B_i observes and receives messages transmitted from the actor A.

Consider how we understand each other in a physical group. The group's norms dictate how one interacts with the other. When one-person speaks, others listen, observe and interpret. These types of interactions help actors to understand each other.

We can represent this understanding of the interaction between sender and receivers within a virtual group. Words, whether spoken or written, become the linguistic vehicle in virtual space. Words indicate the degree of a sender's commitment towards a receiver's expectation, and any associated actions represent translation of these words into reality (Ho, 2008). Different senders (e.g., actor A) will be characterised differently by the receivers (e.g., observers B_i) based on the distinctive patterns of words and actions in their interactions. Receivers, on the other hand, develop feelings of trust toward the sender based on the consistency between his words and actions. This characterisation of the sender by the receivers over time has a direct effect on the building of an ad-hoc consensus among the receivers, and ongoing consistency is a requirement for trust to develop. The presence of words (e.g., commitment to handle deals by the analyst) that are inconsistent with associated actions (failure to submit a final proposal to the sales team by the due date) might reduce consensus among the receivers about the sender's competence to deliver on commitments.

The receivers might have drastically different mental models (i.e., perception) towards a sender during initial

phases of the interaction. The differences among receivers' opinions will tend to merge together and narrow down over time to reflect a general opinion on the sender's trustworthiness. However, for this to happen, a virtual team needs to interact constructively over time. If there is a constant mismatch among the sender's words and actions over time, then trust among the team members will not develop, and consensus towards the sender may be that trustworthiness is low.

A consistent match between the sender's words and actions over time characterises the trustworthiness of the sender. A dramatic shift in the sender's consistency may reduce his perceived trustworthiness. For example, a sudden reallocating of available resources to someone outside the group could be interpreted as an act against the overall interests of the group.

In the previous scenario, if the analyst submits a quote not agreed to by the sales team, his team may deem such act as distinctive behaviour of the analyst and revaluate the scenario to assume there was a push by the sales team to compete with other rivals in the industry. In general, any single act that does not correspond to standard protocols can be recognised as distinctive from normal behaviour. Table 1 describes a set of simple rules regarding comparisons of A's words and actions in conversations that capture the interaction between A and B_i's.

When actor A's actions match to his/her promises (words), actor A is considered by observers B_i as trustworthy. Likewise, when actor A's actions do not match to his/her promises (words), actor A is considered as questionable. If we further analyse this problem, when actor A's actions are less than his/her promises, actor A tends to be interpreted by observers B_i as not trustworthy due to not enough actions matched to those promises. When actor A's actions are more than his/her promises, it does not leave sufficient information or evidence for observers B_i to assess. In this situation, actor A tends to be interpreted as not communicating his/her ideas. The actor A's actions can result in either positive or negative ways depending on the outcome of that action.

3 Trustworthiness

The evaluation of words and actions relate to trustworthiness. In a dynamic interpersonal relationship, trustworthiness is defined as "generalized expectancy, concerning a person's degree of congruence between communicated intentions and behavioural outcomes that are observed and evaluated, which remain reliable, ethical and consistent, and any fluctuation between perceived intentions and actions does not exceed the observer's expectations over time" (Hardin, 2003; Ho, 2009a; Hosmer, 1995; Rotter, 1967); it is an interchangeable concept with dependability or stability. In reality, it is difficult to establish a complete evaluation of trustworthiness toward any social actor from peer networks due to insufficient information and baseline knowledge. There is only a degree of trustworthiness that can be inferred. During social interaction, a finite-state approach based on each transition should be considered rather than analysing the entire communication session as a whole (Ho and Lee, 2012). It is important to note that trustworthiness refers to the inner characteristics of a focal actor, thus some critical contextual information (e.g., situational context, the sending agent's emotional state, the interacting agent's emotional state toward the agent, the baseline or historical knowledge of each agent's response from prior communication, and the character as a whole, etc.) needs to be considered in any trustworthiness assessment situation.

Rather than analysing the direction, frequency, volume, and the duration of conversations, the deciding factor for establishing trustworthiness in a virtual interaction is the textual content, which includes the emotional palette (Ho and Lee, 2012). An effective assessment of a communication can be made when we include all of the content assessment in an approach. In a virtual environment, the sender's intent is communicated by setting forth the agenda (purpose) in the initial stage of the communication. This intent could be transitioned and shifted throughout the ongoing communication. The transition in each agent's information behaviour may imply that the situation is facing a dilemma.

Some of the key characteristics or qualities of a sender, e.g., benevolence, integrity, and competence, can be inferred from the message content (Mayer and Davis, 1999; Mayer et al., 1995). However when an analysis of trustworthiness is conducted, the understanding of actor's above-mentioned internal states (e.g., benevolence, integrity, and competence) should be looked upon separately. Based on the receiving actors' observation, we may be able to infer whether a sender is acting in an ethical manner, or facing a dilemma.

Table 1 A simple illustration of behavioural parameters determining trustworthiness

	Scenarios			
Actor A's behaviour	1) Words = Action	2) Words ≠ Actions	3) Words > Actions	4) Words < Actions
Observer B's perception	Trustworthy	Questionable/not trustworthy	Questionable/not trustworthy	Not enough indication (positive vs. negative)

One of the ways by which we can infer sender disposition is to look for keywords that are characteristic of an attribute. For example, when a sender acknowledges with message patterns that use appreciative, thankful, helping words in a timely manner, these words represent a degree of the kindness of an individual. Kindness is a character trait, and it is likely to be dispositional. Likewise, patterns like 'hmmm', or 'not sure', etc., may represent a delay (a dilemma) in making a choice.

3.1 Internal vs. external factors

Lieberman (1981) classified trustworthiness into two constructs: competence as an external, situational factor – and integrity as an internal, dispositional factor. Competence refers to an actor's confidence level in a particular required skill set. It is one of the positive attributes that can be translated to the building of consensus among receivers. Here, the emphasis is on the ability of the sender (e.g., leadership) to showcase his understanding of the domain, and his capability to address queries posed by the receivers, and by responding to their expectations in a positive sense, he always provides answers with personal responsibility. For example, the sender could use directive tone of voice in his communication with the expression such as 'definitely', 'have completed', 'i can', and 'i will', etc. These words have been predominantly used to demonstrate an actor's competence. The receivers must gauge this attribute from the initial interaction.

4 Attribution mechanisms

Attribution theory concerns causal explanations for human behaviour (Kelley et al., 2003; Kelley and Michela, 1980; Martinko and Thomson, 1998), and reactions to behaviour.

4.1 Kelley's ANOVA model

Kelley et al.'s (1973) ANOVA model is one of the best known among many attribution theories. His theory assumes that people make causal attributions in a rational and logical mode, based on multiple dimensional observations at different times and circumstances. Kelley et al.'s (1973) ANOVA model consists of three causal factors (i.e., persons, stimuli, and times) and three types of information (i.e., consensus information, consistency information, and distinctiveness information) that people use to make attribution decisions to one's behaviour. *Consensus* refers to the extent to which other persons respond to stimulus in the same way. *Consistency* refers the extent to which a person's response to the stimulus is consistent over time. *Distinctiveness* refers to the extent to which the person's response to the stimulus is distinctive to his or her responses to other stimuli. Kelley et al. (1973) identified eight possible patterns of high versus low regarding these three types of information: consensus, consistency, and distinctiveness. The high, high, high pattern can be attributed to the stimulus (external cause); the low, high, low pattern can be attributed to person (internal cause); and the low, low, high pattern can be attributed to the circumstance (external cause). Although Kelley et al.'s (1973) ANOVA model has been tested in several studies (e.g., McArthur, 1972), these three types of information may be difficult to collect in order to make logical causal attributions.

Kelley et al.'s (1973) ANOVA model is ideal, logical and sound, but in reality, people may not make attributions in a rational way without pre-existing suppositions about cause and effect (Kelley and Michela, 1980). In studying leadership, for example, people generally believe that the success of any unspecified person is attributed more to factors within the person (e.g., ability, effort) than external factors (e.g., an opportunity, networked relationship). Moreover, an actor's self-centred motive tends to favour him- or her-self and thus introduces bias into the attribution process. For example, actors, due to motivations for self-enhancement and self-protection, tend to attribute internal causality toward themselves for any positive behavioural outcome, and attribute external causality toward others for any negative behavioural outcome. Actors are motivated to make attributions that present themselves in a favourable way to observers.

Table 2 Patterns of information types in Kelley ANOVA model

Information type	*Pattern*		
Consensus	High	Low	Low
Consistency	High	High	Low
Distinctiveness	High	Low	High
Attribution	External causes -> Stimulus	Internal causes -> Disposition	External causes -> Circumstances

Source: Kelley et al. (1973)

4.2 Internal vs. external causality

Kelley et al. (1973, 2003) and Kelley and Michela (1980) suggests that group consensus among the receivers is arrived at based on the evaluation of the consistency between the sender's words and action over time, and also based on any distinctive behaviour within a collective baseline profile. However, receivers or observers always evaluate a sender's actions, and assign an either internal or external cause to anomalous behaviour. When external causality is perceived, it means that the receivers do not attribute to the sender responsibility for an action. When internal causality is perceived, it means the receivers may assign the sender personal responsibility for an action.

In our example of a financial analyst, suppose that the sales team pushes the analyst for a lower bid to win against a rival competitor. Now, assume that the analyst does try to give a lower quote (words) but failed to complete for the transaction (actions). On the first such occasion, the inconsistency between words and actions might be attributed to external causality. Although the analyst's failure (action) is distinctive, it would not be attributed with nefarious intention. As a result, the analyst is not held responsible for this action.

On the contrary, suppose the analyst is persuaded by a client to deceive by forwarding a lesser quote that is not approved by the management. This analyst's betrayal action would increase the chances of distinctive behaviour by significant degrees. It may be likely that his team-mates would notice small inconsistencies between the analyst's words and actions. His words and actions fail to be consistent. We assert that any single instance of a distinctive behaviour is likely to be evaluated by the group and the stimulus could be attributed to be internal (dispositional) or external (boss tells him to do it without telling the group, or he learns something that influences his decision, or any number of similar circumstances could exist). Since the interaction of the analyst with the team and the management is always a close association, the group shares a common mental template toward the analyst's historical behaviour. The group may attribute the analyst as someone with high integrity since there was never a distinctive behaviour observed according to their historical observations. The distinctive behaviour by the analyst may impact the group's ability to maintain a consensus on his trustworthiness. As the group is highly sensitised, they may notice inconsistencies between the analyst's words and actions, and the distinctive behaviour, which is different, compared to their historical observations of the analyst. When the analyst's actions change the original perception of the team members, his integrity could be re-evaluated, and his actions could be attributed to internal causality. In other words, the analyst will be viewed as responsible for the fault of this action.

5 Dyadic attribution mechanism

Ho and Benbasat (2014) propose a theory of trustworthiness attribution for identifying potential insider threats in cyber organisation. Insider threat refers to situations where a critical actor of a virtual team betrays the team members in an illegal or unethical manner. Betrayal is a cognitive construct that indicates a violation of interpersonal trust (Elangovan and Shapiro, 1998). In Ho and Benbasat's (2014) study, trustworthiness acts as a window into an actor's dispositional state of ethical standards, which can serve as a precursor to the likeliness of betrayal. Ho and Benbasat's (2014) theory proposes that through internal causality over the focal actor's trustworthiness in online interaction, the downward shift of actor's likelihood in betraying against the virtual team or organisation can be observed through the actor's explicit information and communication behaviour by those who are in close relationship with the actor. Moreover, a group tends to have less agreement toward a social actor's lack of consistency between words and actions; and, the distinctiveness of his or her information behaviour is high when the tendency of this focal actor's trustworthiness is found to be relatively low. In a way, the demonstration of an actor's information behaviour (based on these three information types) may infer that an actor's disposition has a tendency for a betrayal. On the other hand, a group tends to move towards a higher consensus based on a social actor's high consistency between words and actions. The distinctiveness of an actor's information behaviour is low when the tendency of this actor's trustworthiness is found to be relatively high. In this scenario, we may say that the actor may have high loyalty, resulting in low tendency or reaction to betrayal.

5.1 Consistency of words and actions

In order to assess an actor's attributes within a given context, it is important to interpret the content of the messages exchanged. Consistency in the virtual environment interaction is defined as a sequence of communication in which the sender's words and actions complement each other throughout. Inconsistency of a sender's words and actions could occur during intermittent context shift scenarios. It refers to the consistency of the sending actor, irrespective of what the receivers perceive. The initial communication assessment from the receivers is taken into account as baseline observations. When evaluating the sender's communication patterns, it is possible for receivers to experience a change in the communication pattern during the course of communication whenever there is a shift in the context content.

5.2 Group consensus

Group consensus can only develop within a team that has interacted for a fair amount of time under a certain context. Arriving at consensus is difficult under scenarios where there may not be enough interaction. Typically, the group is led by a social actor who facilitates discussion among others receivers. Receivers share a common opinion about the social actor (sender) though interactions over time. By analysing the content of the messages, we may infer on the consensus built toward the sender by the others.

Unlike consistency, the agreement between the sender and the receivers plays a vital role in building the consensus. Consensus is a degree of collective measure of the agreement among the receivers toward the sender. Frequent disagreement among the receivers toward the sender hinders the effort to build consensus.

5.3 Distinctive behaviour

Ethical behaviour of a social actor can be sensitive and complicated to infer. The others can predict this only after having known the actor over time. They must be capable to detect any kind of abnormality in the senders' messages. They must be able to see a clear deviation from the expected response from the sender. This can be even more evident when an actor changes communication patterns across similar situations from his normal and regular behaviours. It also refers to the actor acting differently from other actors in the same or similar situation. For example, if the sender, who is supposed to act as a facilitator in a discussion, decides to dictate terms and represents disagreement throughout, such behaviour is different from how facilitators in general would act. This is categorised as a distinctive behaviour of the actor (the sender). He or she does not act in accordance to the baseline organisational principles and policies. As a result, his or her behaviour may represent an act of personal gain rather than the collective interest.

6 ABM of the dyadic attribution mechanism

ABM provides a paradigm that allows us to model a social actor's behavioural changes at an individual level. An ABM can capture the emergent phenomena, stochastic behaviour changes, heterogonous population, and offers a natural and flexible way to describe a system. It therefore has been used in numerous fields to study, for example, infectious disease transmission, new product adoption, etc. Agents can refer to any real world entities, such as person, projects, companies, animals, products, etc. An agent can be designed to represent its behaviour, attributes, their reactions to stimuli/environment, and interaction with others, e.g., agents, environment, etc. The physical and emotional influences on each entity, along with its interactions within the social environment can be quantitatively captured in an ABM. In our study, the internal state – and behaviour of an agent – will be represented using variables and the states with the AnyLogic™ (XJ Technologies, Russia) simulation environment.

Bonabeau (2002) discusses this modelling technique with a number of real life examples that employ agent-based models. Urban and Schmidt (2001) studied modelling of complex internal states and the interaction between physical and psychical processes among agents. The physic, emotion, cognition, social status (PECS) reference model was created to provide concepts for the construction of such human-like agents. The architecture of PECS contains three different layers. The first layer contains components for sensors, which are responsible for processing input data. The second layer processes the internal state of an entity. The third layer comprises the actor's outward behaviour and actor components, which evaluate the agent's behaviour and execution of actions. In ABM, the input information of visual and audio was considered in a virtual environment. Macy and Willer (2002) studied autonomous decision makers interacting with other local agents for key assumptions in ABM. Macy and Willer (2002) emphasised that no system can be built with a top down approach as one globally integrated entity. A model must be implemented with a bottom up approach, recognising smaller components that make up the system and their interactions. Agents are interdependent of one another as they engage in the processes, which might influence other agents. This interdependency can be indirect. For example, when an agent's behaviour changes it might impact the decisions of other agents, which may cause a change in the environment.

6.1 Rules construction

In our study, an agent-based model is set up to capture the process of the communication/dual communication channel down to an individual level between the sender (A) and the receivers (B_i) quantitatively. Simply put, A or B_i can send messages to each other during the communication, and there is a threshold set for whether one trusts the other or not, depending on the context of the text. The historical trustworthiness between A and B_i is computed by correlation function that captures the consistency of previous communication/behaviour between A and B_i – and current interactions – in order to detect potentially suspicious behaviour of A. The threshold ~ [–1, 1], therefore, is set up to trigger this detection. The modelling tool called AnyLogic™ will be used to simulate different scenarios as described above.

Below we define some rules and logics of how to model the interactions between A and B_i, and how B_i evaluates A's behaviour. This evaluation is based on B_i's individual opinions, assuming that A's action (Table 1) is identical to all B_i, but B_i's observation/perception towards A can be heterogeneous.

1 Actor A will 'broadcast' his action to all B_i, which means each B_i sees identical action of A.

2 *Trustworthiness*: Actor A will have individual communication with B_i at each time milestone, and B_i will evaluate A's words (Table 1) to calculate a quantitative value about his/her opinion of A. We assume that it is a heterogonous group of B_i (different sets of mental models), and therefore they will have different opinions and evaluation over A's action. If we quantify agent B_i's evaluation (1 means trustworthiness and –1 means betrayal), the evaluation ranking can range anywhere between –1 and 1.

3 We use the emotions of B_i as a measure of performance evaluation (PE). That is, if the emotional state of B_i is positive (e.g., happy), it is interpreted as an act of trust given by B_i. If the emotional state of B_i is negative (e.g., confused, angry), it is interpreted as an act of distrust by B_i. There are however exceptional cases. The expression of 'worry', for example, depicts a neutral state of a person in terms of perceived trustworthiness. A social actor B_i can share the emotion of worry, and such emotion does not necessarily reflect concerns for ethical dilemma.

4 This PE value will also depend on B_i's historic opinion about A's performance at previous milestones, (A's performance at current time step under the consideration of A's historical performance). It will be dynamically update at each milestone until the end of the simulation. We will get the final score of A updated from B_i, which is the average of all performance scores of A from all B_i's evaluation.

5 *Stochasticity of the simulation* is captured by the occurrence of uncontrollable situations, which may affect the PE of actor A. The direction of the communication, the frequency, the duration, the content and the volume must be all considered in setting the scenario. In addition to the situational context, the individuals' emotional state, comfort level between the sender and the receiver, knowledge of the receivers' responses from prior communication, and the character of each actor as a whole are more sophisticated in establishing trustworthiness.

6 Consistency of A's words and actions: an integer variable is created for any iteration of the simulation; the value is incremented by 1. If the integer value is 1, there is a match between A's words and action, if not it will be decremented by 1. A match between a sender's words and actions is a representation of the consistency of the sender. A large positive score determines the consistency of the sender A over time.

7 Distinctive behaviour of A: Suppose consistent ratings of A's PE, whether positive or negative, can be found in B_is' evaluation toward A. If a high percentage of the total positive count results as norms of B_is evaluation of A (PE_A), a few opposite, meaning negative, counts can refer to as A's distinctive behaviour in different time, across similar situations. If high percentage of the total negative count is found in B_is evaluation of A (PE_A), a few opposite, meaning positive, counts can refer to as A's distinctive behaviour in different time, across similar situations.

8 Consensus among B_is: An environment variable called consensus can be set up to capture inputs from B_is. The inputs from each receiver B_i are aggregated and calculated. The group consensus can be determined by comparing the total positive count for 1 and total negative count for –1 toward actor A. If the total positive count for 1 is more than the total negative count for –1 from B_is, we assume that group B_is has higher consensus toward A and shares a belief that A is trustworthy. If the total positive count for 1 is less than the total negative count for –1 from B_is, we assume that group B_is has lower consensus toward A and believes that A is untrustworthy. If the total positive count for 1 equals or not much more or less than the total negative count for –1 from B_is toward A, we assume that group B_is can not reach consensus toward A.

9 Ethical dilemma: Every sender A will face two choices: one choice would benefit the individual self, and the other would benefit the group. When the benefit of the individual fits within the benefit of the larger self, these benefits do not contradict each other. In this condition, the sender A could choose the benefit to the group while the individual would also be benefit. These choices do not contradict one another. Regardless, the sender A's choice is considered ethical. When the benefit of the individual is in direct opposition to the benefit of the group, the sender A will face two choices, and could choose either one. This situation when the actor is presented with two ethical choices is considered the moment of ethical dilemma. If the sender A chooses benefits for the group over the benefit of the self gain when there is a conflict of interests involved (meaning when personal benefits are in conflict with the group benefits), the sender A's choice is considered ethical. By contrast, if the sender A chooses the benefits to the self gain over the benefits to the larger self during the conflict of interests situation, the sender A's choice would be considered unethical.

6.2 An example

We have converted data from an online game simulation conducted in Ho and Warkentin's (2013) study into an Excel spreadsheet. There are four teams (teams alligator, buffalo, crocodile, dragon) in this experiment. Team-leaders as targets work together with their team members in each online collaborative group work. Team-leaders alligator and buffalo were not influenced with the ethical dilemma presented in the form of incentive money while team-leaders crocodile and dragon faced ethical dilemma situation where they betrayed their team members.

Each agent's conversation with each other is processed at the sentence level with sentiment analysis. We built an agent-based model using AnyLogic™ to take emotional inputs of each agent in an excel spreadsheet. The emotions of each agent are processed and displayed in text, and the interactions among all agents in a community are displayed chronologically (Ho and Lee, 2012).

These Excel spreadsheets are connected into the AnyLogic™ environment using the function, *connectivity property*. We identified the number of the members in a group ($i \leq 7$). Each spreadsheet represents one person (or, member) in a group. For each agent (p), we use variable j to represent the total records of the conversations spoken by this agent corresponding to the i value where it indicates the number of agents in a chat room. Below is the code used to create the agents.

```
Person p;
for(int i=1;i<=7;i++){

    Int k = excelFile.getLastRowNum(i);
  p=add_people();
```

In our agent-based model, we used *date* as a key to reflect the emotion as well as the trustworthiness of each agent in the game. Below is the code used to parse through the records (dates).

```
for(int j=2;j<=k;j++){
  Date d = null;
  String sd =
  excelFile.getCellStringValue(i,j,1);
  java.text.SimpleDateFormat sdf = new
  java.text.SimpleDateFormat("MM/dd/yy
  hh:mm aaa", Locale.US);
    try{
      d = sdf.parse(sd);
    } catch(Exception e){};
```

Then we mapped the corresponding emotions of these created agents into the model as per the *date* and *time* from the Excel spreadsheets, so that the emotions are displayed according to the conversation or chat that occurred. Here the emotions are displayed as text (Figure 1). To map the emotions according to the date, we have made the date the key, and emotion as a value and created a linkedhashmap (data structure in java) by name emotionsByDates. Similarly we have created a linkedhashmap (data structure in java) by name trustBydates for mapping the trustworthiness values.

```
p.emotionsByDates.put( d,
excelFile.getCellStringValue(i,j,3) );
```

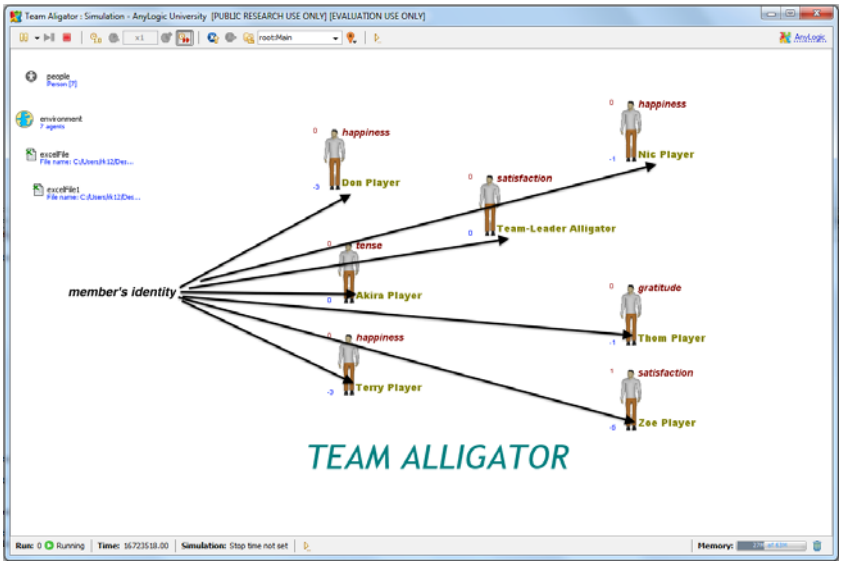
We have also mapped the trustworthiness values again according to the *date* and *time.*

```
p.trustBydates.put(
d, excelFile.getCellNumericValue(i,j,4)
);
```

When emotions and trustworthiness values are read, we use the following code to show the actions being triggered. The total count of trustworthiness value is kept.

```
if( emotionsByDates.get( date() ) != null
&& !emotionsByDates.get( date() ).equals(
emotion ) )
{
    emotion = emotionsByDates.get( date()
    );
    System.out.println(emotion);
    trust = trustBydates.get( date() );
    count=count+trust;/*keeps track of
    total trustworthiness*/
}
```

Figure 1 Online interactions for alligator virtual team (see online version for colours)



We use this total trustworthiness value at the end of the simulation to see which team has more confidence or trust in their team leader. Positive values show their positive emotions and positive trustworthiness evaluation over their team-leader while negative values show team members' negative emotions and negative trustworthiness evaluation toward their team-leader.

After creating the agents for mapping their emotions and trustworthiness values into the model we used the simulation experiment component in the model and simulated the emotions and trustworthiness of the players as shown in Figures 1 to 4. In these figures, agents' identities, roles, their emotions, and team members' emotions and trustworthiness attribution values toward their team-leaders are represented around each agent.

This ABM approach allows the research to automatically visualise the trustworthiness of the targets, team-leaders from the perspectives of their team members.

7 Conclusions, contribution and future work

This paper describes a set of rules as basic logic for modelling deceptive behaviour in a virtual forum. We model and visualise the trustworthiness of a social actor to automatically signal potential for betrayal through sentiment analysis at the sentence level in insider threat situation. This concept can be further extended to mimic real-time, real-world computer-mediated communications by processing textual data to infer an actor's trustworthiness.

Figure 2 Online interactions for buffalo virtual team (see online version for colours)

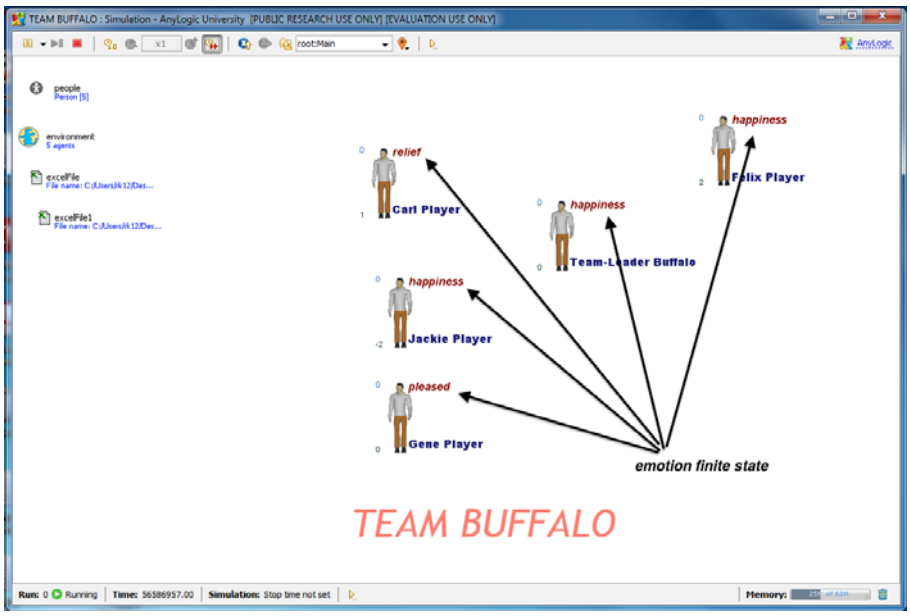


Figure 3 Online interactions for crocodile virtual team (see online version for colours)

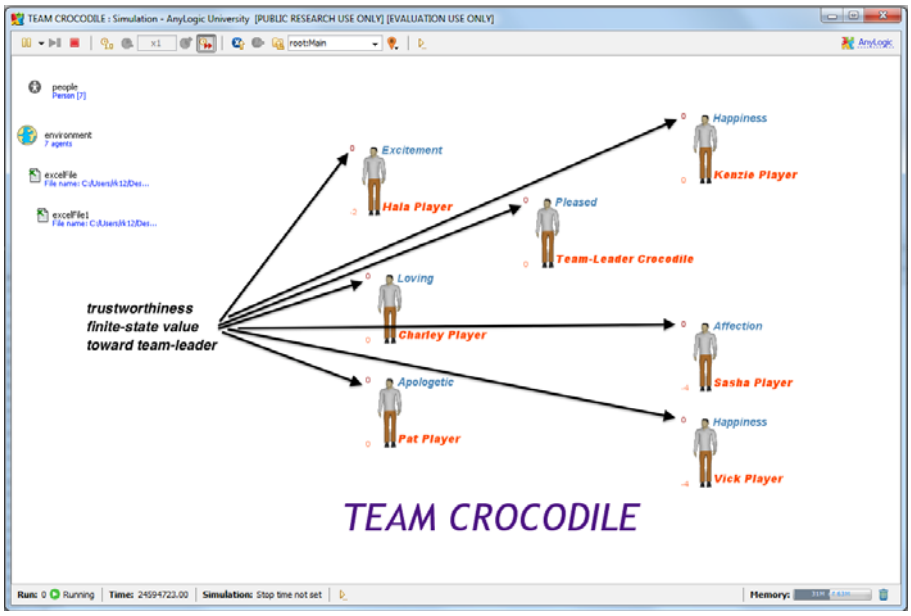
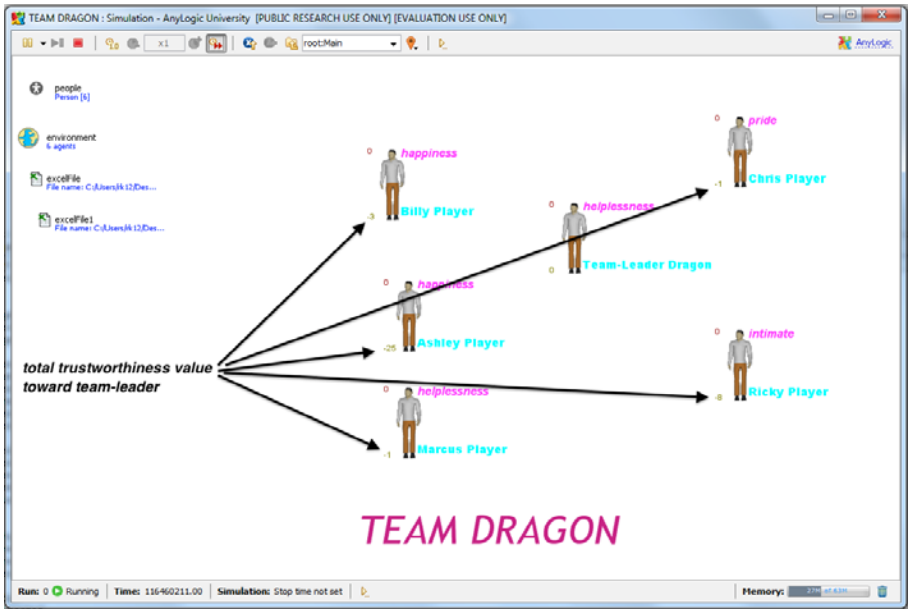


Figure 4 Online interactions for dragon virtual team (see online version for colours)



Inevitably, the model can only be useful when it is constructed with the right level of the descriptions and measurement schemas of each agent's attributes, behaviour, and their interacting environment, which ultimately requires a large amount of data. This simple logic represents how science can be encapsulated in art, which plays a critical role in the model estimates of cases and the dynamics of evaluation. The parameters and rules constructed in ABM are often justifiable. However, simple logic can turn out to be insufficient and rather complicated during implementation and deployment.

This study provides an example of the potential use of an ABM to study a dyadic attribution mechanism within covert human interactions, and shows the importance of assumptions about human interaction. The simulation output can provide quantitative forecasting and insight into possible results. These results may not be a consistently accurate prediction. ABM can illustrate varied behaviour of a social actor in a systematic way that could be rarely found in the real world. ABM is a powerful tool for understanding the behaviour of complex systems, especially when it involves the heterogeneity of human agents and their interactions (e.g., through networks). It will be necessary to conduct multiple runs of the simulation by varying the initial conditions and parameters to ultimately assess the validity of the results.

Acknowledgements

The authors would like to thank Kun Hu from IBM Research, Shuheng Wu, Jonathan Hollister, Dhaval Kashyap and Miikael Lehto from Florida State University, and Pavel Lebedev from AnyLogic for their contribution and support of our research discussion. The first author wishes to also thank Conrad Metcalfe for his editing assistance.

References

Bonabeau, E. (2002) 'Agent-based modeling: methods and techniques for simulating human systems', in *Proceedings of the National Academy of Sciences of the United States of America*, PNAS, pp.7280–7287.

Elangovan, A.R. and Shapiro, D.L. (1998) 'Betrayal of trust in organizations', *The Academy of Management Review*, Vol. 23, No. 3, pp.547–566.

Hardin, R. (2003) 'Gaming trust', in E. Ostrom and J. Walker (Eds.): *Trust and Reciprocity: Interdisciplinary Lessons from Experimental Research*, pp.80–101, Russell Sage Foundation, New York.

Ho, S.M. (2008) 'Attribution-based anomaly detection: trustworthiness in an online community', in H. Liu, J.J. Salerno and M.J. Young (Eds.): *Social Computing, Behavioral Modeling, and Prediction*, pp.129–140, Springer, Tempe, AZ.

Ho, S.M. (2009a) 'Behavioral anomaly detection: a socio-technical study of trustworthiness in virtual organizations', in School of Information Studies, Syracuse University, Syracuse, pp.1–437.

Ho, S.M. (2009b) 'A socio-technical approach to understanding perceptions of trustworthiness in virtual organizations', in H. Liu, J.J. Salerno and M.J. Young (Eds.): *Social Computing, Behavioral Modeling, and Prediction*, pp.113–122, Springer, Tempe, AZ.

Ho, S.M. and Benbasat, I. (2014) 'Dyadic attribution model: a mechanism to assess trustworthiness in virtual organizations', *Journal of American Society for Information Science and Technology*, forthcoming.

Ho, S.M. and Lee, H. (2012) 'A thief among us: the use of finite-state machines to dissect virtual betrayal in computer-mediated communications', *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA): Special Issue of Frontiers in Insider Threats and Data Leakage Prevention*, Vol. 3, Nos. 1/2, pp.82–98.

Ho, S.M. and Warkentin, M. (2013) 'Simulating insider threat with the leader's dilemma game: an experiment protocol', *Information & Management* (under review).

Hosmer, L.T. (1995) 'Trust: the connecting link between organizational theory and philosophical ethics', *Academy of Management Review*, Vol. 20, No. 2, pp.379–403.

Kelley, H.H. and Michela, J.L. (1980) 'Attribution theory and research', *Annual Review of Psychology*, Vol. 31, pp.457–501.

Kelley, H.H., Holmes, J.G., Kerr, N.L., Reis, H.T., Rusbult, C.E. and Van Lange, P.A.M. (1973) 'The process of causal attribution', *American Psychology*, Vol. 28, No. 2, pp.107–128.

Kelley, H.H., Holmes, J.G., Kerr, N.L., Reis, H.T., Rusbult, C.E. and Van Lange, P.A.M. (2003) *An Atlas of Interpersonal Situations*, Cambridge University Press, New York, NY.

Lieberman, J.K. (1981) *The Litigious Society*, Basic Books, New York, NY.

Macy, M.W. and Willer, R. (2002) 'From factors to actors: computational sociology and agent-based modeling', *Annual Review of Sociology*, Vol. 28, pp.143–166.

Martinko, M.J. and Thomson, N.F. (1998) 'A synthesis and extension of the Weiner and Kelley attribution models', *Basic and Applied Social Psychology*, Vol. 20, No. 4, pp.271–284.

Mayer, R.C. and Davis, J.H. (1999) 'The effect of the performance appraisal system on trust for management: a field quasi-experiment', *Journal of Applied Psychology*, Vol. 84, No. 1, pp.123–136.

Mayer, R.C., Davis, J.H. and Schoorman, F.D. (1995) 'An integrative model of organizational trust', *Academy of Management Review*, Vol. 20, No. 3, pp.709–734.

McArthur, L.A. (1972) 'The how and what of why: some determinants and consequences of causal attributions', *Journal of Personality and Social Psychology*, Vol. 22, No. 2, pp.171–193.

Rotter, J.B. (1967) 'A new scale for the measurement of interpersonal trust', *Journal of Personality*, Vol. 35, No. 4, pp.651–665.

Urban, C. and Schmidt, B. (2001) *PECS – Agent-based Modeling of Human Behavior*, in AAAI Technical Report FS-01-022001, AAAI, pp.1–6.