2017

Trustworthiness Attribution: Inquiry into Insider Threat Detection

Shuyuan Mary Ho, Michelle Kaarst-Brown and Izak Benbasat

This article was publihsed in the Journal of the Association for Information Science and Technology. The publisher's version is available at https://doi.org/10.1002/asi.23938.

TRUSTWORTHINESS ATTRIBUTION: INQUIRY INTO INSIDER THREAT DETECTION

Shuyuan Mary Ho[*]
Assistant Professor
School of Information
Florida State University
142 Collegiate Loop
Tallahassee, FL 32306-2100
smho@fsu.edu
(850) 645-0406

Michelle Kaarst-Brown
Associate Professor
School of Information Studies
Syracuse University
218 Hinds Hall
Syracuse, NY 13244-1190
mlbrow03@syr.edu
(315) 443-1892

Izak Benbasat
Professor Emeritus
Management Information Systems
Sauder School of Business
University of British Colombia
Vancouver, BC Canada V6T 1Z2
izak.benbasat@sauder.ubc.ca
(604) 822-8396

ABSTRACT

Insider threat is a "wicked" contemporary organizational problem. It poses significant threats to organizational operations and information security. This paper reviews insider threat research and outlines key propositions to conceptualize the interpretation of dynamic human information behavior in an organizational setting, which represent an integration of trustworthiness and human sensors' attribution in close relationships. These propositions posit that when a focal individual violates integrity-based trust, the group can collectively attribute a shift in trustworthiness, triggering a natural peer attribution process that assigns cause to observed behavior. Group communication can thus reflect subtle changes in a focal individual's perceived trustworthiness. The ability to understand group-based computer-mediated communication patterns over time may become essential in safeguarding information assets and the "digital well-being" of today's organizations. This paper contributes a novel theoretical lens to examine dynamic insights on insider threat detection.

Keywords

Trustworthiness attribution; sociotechnical systems; insider threat; betrayal; trusted human computer interaction; information behavior; computer-mediated deception.

Introduction

"Over the past century, the most damaging U.S. counterintelligence failures were perpetrated by a trusted insider with ulterior motives. In each case, the compromised individual exhibited the identifiable signs of a traitor—but the signs

---

[*] Corresponding author.

went unreported for years due to the unwillingness or inability of colleagues to accept the possibility of treason" (ODNI, 2014)[1].

One of the greatest threats to organizational security is exposure of information (e.g., intellectual property) through espionage. Cyber espionage is on the rise (Verizon, 2017, p. 10), and 90% of cyber espionage breaches are designed to capture trade secrets and proprietary information (Verizon, 2016, p. 54). Insider threats unfortunately pose an increasingly significant problem for trusted interactions in both physical and virtual organizations. An extreme example is the infamous case of Robert Hanssen, a U.S. counterintelligence agent, who (ab)used his privileges and position of trust to sell highly classified national security materials to the KGB/SVR in Soviet Union/Russia *over a period of 15 years* in exchange for personal financial gains. This case of espionage by a *trusted* and *authorized* insider shows how insider malfeasance can adversely affect an organization (FBI Press Release, 2001), and illustrates the importance—as well as the challenges—of uncovering deceptive behavior within an organization.

One universal fact is that most organizations, whether public or private sector, must rely on their employees, i.e., key "insiders," to attain performance and productivity. These key individuals have access to information that could result in reputational, financial, or productivity losses if misused. *Insider threat* is a reference to situations in which a "focal actor"—someone with authorized access—inflicts damage to their own organization by behaving against the interests of the organization (i.e., betraying), generally in an illegal and/or unethical manner (Ho & Benbasat, 2014). Ultimately,

[1] ODNI stands for the Office of the Director of National Intelligence, established in 2004 as part of the Intelligence Reform and Terrorism Prevention Act.

insider threat always involves some aspect of betrayal, which is an intentional act of trust violation against the interest of another party, such as to "expose a secret" without the consent, agreement or authorization. While there are many types of non-malicious (well-intentioned or negligent) insiders who might inadvertently betray the organization, there are also those who do so maliciously, with deliberate intent to harm the organization for some benefit.

The equally infamous case of Edward Snowden, a former National Security Agency (NSA) contractor, exemplifies well-intentioned betrayal. Regardless of whether Snowden's motivation for leaking classified government information was to alert the public about the surveillance state of the government, or was based on his interest in freedom and privacy, the result significantly impacted U.S. intelligence operations and reputation (Times, 2014). Likewise, Chelsea Elizabeth Manning (formerly Bradley Edward Manning) who posted U.S. Army's classified documents to WikiLeaks, was convicted for violating the Espionage Act (18 U.S.C. § 792 et seq.) (Tate, 2013). Regardless of their underlying motivation, these clearly amount to acts of betrayal against an organization.

Insider threats are certainly not limited to governmental organizations. As the trigger for insider threat attacks is typically financial, or for purposes of corporate espionage, the private sector also faces serious financial, data, and reputational loses. The Verizon (2017) data breach investigation report (DBIR) indicated that 94% of breaches had a financial or espionage motive (pp. 3 & 5), and breaches involving insiders had increased 12% since last year (2016). Unprecedented information access, availability, and

connectivity for key employees have made organizations increasingly susceptible to insider threats—especially from disgruntled or ex-employees. Bolstering security against the weakest link, *the human factor*, is thus critical in an organization's information security defense (Mitnick & Simon, 2002).

This paper provides inquiries into insider threat research, and moreover contributes to gaps in existing research by arguing for a new perspective of trustworthiness attribution to examine insider threat situations through collective group attribution that reflects the perceived trustworthiness of a focal actor in group communication.

Inquiry into Insider Threat Research

Information, as an artifact of an organization's asset and knowledge, is subject to the omnipresent threat of unauthorized duplication, modification, and disclosure by insiders. We can lock down information like we protect gold in a vault, but once classified information is revealed, copied and disseminated, it is forever compromised and can never be made secret again. Thus, a more sophisticated approach is required to prevent and to protect against information breaches from insiders.

Traditional approaches investigate on an already identified breach, or create policies that deter potential perpetrators. Unfortunately, the current state-of-the-art technology has limitations in its ability to infer complex patterns of human behavioral anomalies. Expecting "whistleblowing" by colleagues to counter insider threat has also become increasingly unlikely. Perpetrators' motives are impossible to identify with certainty before incidents happen. The social norms in our democratic society encourage

information access, discourage corporate monitoring, and value personal privacy, which has made surveillance awkward and frequently unethical. Below we discuss inquiries from both technological and behavioral domains that will lead to a new proactive perspective on insider threat detection—whereby future system design can be informed by psycho/social-behavioral theories.

Insights from Technological/Computational Literature

From computational literature, we first identify that due to the sophistication and increased use of information communication technology (ICT), cloud environments have enabled a greater magnitude of damage, and made infrastructure even more vulnerable to insider threat activities (Patel *et al.*, 2013). The case with Snowden illegally copying 1.7 million documents from the N.S.A.'s internal SharePoint systems (a private cloud) without authorization is an illustration (Toxen, 2014). Second, current computational approaches require large "appropriate" and "complete" datasets to train the system to analyze and compare patterns of behavior with pre-defined rules (Debar *et al.*, 1999). However, it is unrealistic to assume a complete dataset on targeted human behaviors. Inappropriate or incomplete datasets limit analysis, producing imprecise and noisy results. Third, although an insider's behavior may be camouflaged as part of normal work activities, unsophisticated technology of "misuse detection" that provides a weak indication of intent, still cannot protect against insider threat activities. Fourth, intrusion detection and prevention systems (IDS/IPS) are designed to analyze network-based attacks. As such, these systems tend to produce false alarms, irrelevant to infer human intent. The ineffectiveness of current technological solutions points to an alternative way

of adopting an interdisciplinary approach, that combines social and technological methods (Willison & Warkentin, 2013).

Insights from Human Behavior Literature

The human behavior literature on insider threat studies has primarily focused on retrospective description, or meta-analysis of past crimes. Moore *et al.* (2009) classified four types of insider threat situations: IT sabotage, theft or modification of information for financial gain (fraud), theft of intellectual property (IP) for business advantage, and corporate or government espionage for other reasons, and argued that each requires different technical deterrents. Although information policy can provide preventive measures, the impact of punishment as a deterrent to stop individuals from unauthorized attempts is limited.

Not all insider security breaches or information misuse can be considered malicious. Some are "well meaning" or unintentional, such as the debatable cases of Snowden (Times, 2014) or Manning (Tate, 2013). However, they do not share a common profile or motivation (Wall, 2012). Several studies concluded that future research on insider threats needs to include understanding about the thought processes of potential offenders (Willison & Warkentin, 2013). Detecting cues to behavioral changes (whatever the cause) as early as possible becomes increasingly important. However, most of these studies focus on individual perpetrators, with no attention to understanding group interactions surrounding actual or potential incidents.

In spite of the fact that extant research has not identified any effective, proactive approach to detecting insider threat while it is occurring, there are a few studies that

provide valuable insights on the phenomenon. First, although the intentions of the perpetrators (i.e., betrayers) cannot be predicted because they vary greatly, they tend to provide communication cues of their intent (Ho *et al.*, 2016). Second, insider betrayal tends to represent *socio-psychological behavioral problems* that are fundamentally difficult to detect in advance (Liang *et al.*, 2016). The ability to obtain measures of *a person's baseline behavior* is important in efforts to predict future malfeasance (Debar *et al.*, 1999). Third, understanding of insider activities can aid in earlier detection and possibly reduce future incidents of insider threat. Fourth, understanding group interactions could help with early threat detection (Ho *et al.*, in press). That is, during group interaction, people are able to draw metacognitive inferences from each other's memory performance (Smith & Schwarz, 2016). In close proximity, collective group cognition could make it possible to sense and reflect suspicious practices when enabled by distributed metacognition and shared memory (Schwarz, 2015). Lastly, collective interactions in communications could increase observational opportunities to identify potential anomalous behaviors. Empirical research indicates that ethical dilemmas can be attributed collectively by group members during interactive online communication (Ho *et al.*, in press). Communication artifacts, such as blog posts, texts, and emails produced during group activities can be codified to assess shifts in the trustworthiness of a focal actor from the perspectives of a group's collective assessment (Ho & Warkentin, 2017).

Significance of New Perspectives

Insider betrayal poses significant threats to trusted interactions and collaboration within organizations. The problem of insider betrayal becomes more complex when

computer-mediated technologies and the cloud infrastructure enable information access and facilitate information sharing and communication. The ability to understand and detect how an individual's communication patterns morph during deceptive activities—and how anomalous behavior is attributed over time by group members—is essential in safeguarding information assets and the "digital well-being" of organizations. A next-generation framework for detecting insider betrayal should conceptualize the dynamic causal relationships aimed at behavioral changes during human interaction from the collective perspectives of group attribution. This framework contributes an alternative theoretical lens to evaluate insiders' trustworthiness and detect behavioral changes before the occurrence of an anomaly that impacts the operations of an organization. To summarize, new perspectives on trustworthiness attribution can:

1. Provide early warning insights of insider anomalous activity from socio-psychological perspectives (e.g., Wall, 2012).
2. Utilize humans' distributed metacognitive abilities (Schwarz, 2015; Smith & Schwarz, 2016) to correlate complex observations in order to sense and filter subtle cues of a focal actor's trustworthiness in group communication (e.g., Ho & Benbasat, 2014).
3. Incorporate collective group intelligence to process dynamic social interaction and communication rather than focusing on individual activities (e.g., Ho *et al.*, in press).
4. Correlate observations and data collected from social media and computer-mediated communication environments (e.g., Ho *et al.*, 2016).
5. Accommodate a panoramic view of the dynamics of insider activities (e.g., Ho & Warkentin, 2017).

The following section introduces our theoretical stance of trustworthiness attribution. We first conceptualize the insider threat scenario, define the research constructs, and then postulate 15 propositions based on the merger of trustworthiness and attribution theories.

Trustworthiness Attribution

Prior research suggests that a perpetrators intent is usually embedded in either face-to-face relationships, or online communication (Ho & Benbasat, 2014). From a theoretical perspective, the challenge of capturing a perpetrator's intent and behavioral changes is to move beyond "whistleblowing," to early detection based on empirical evidence. Ho and Benbasat's (2014) dyadic attribution model provides a theoretical foundation for analyzing the causal relationships underlying behavioral observations, as indicators of trustworthiness. Specifically, the dyadic attribution model can be used to explain the influence of an actor's information behavior (e.g., language-action cues) based on the observer's perception of the actor's trustworthiness (or lack thereof). Social actors' use of language in social interactions can reveal symbolic clues to behavioral trends, and serve as an observational mechanism to signal shifts in trustworthiness. Based on this, we argue that one way to detect insider deceptive behavior is by assessing shifts in a group's perception when reflecting a focal actor's "trustworthiness" through the analysis of group communication (Ho *et al.*, in press). Unusual or unexpected changes (anomalies) in an individual's perceived trustworthiness as observed by associated work groups may provide early clues to insider threat activity.

Two considerations are required to conceptualize trustworthiness attribution: 1) capturing behavioral evidence in communication artifacts that can objectively signal a change (i.e., downward shift) in a focal actor's behavior, and 2) providing a framework that explains the causal relationships behind a group's attributions of a focal actor's trustworthiness. Below, we first conceptualize on the trustworthiness of social actors.

Then we review attribution theory, discussing behavioral evidence as observed and attributed by others in close relationships, defining research constructs, and explaining a merger of these two theories to provide a detection framework for insider threat.

Conceptualization

The trustworthiness of an insider through his/her communication (i.e., language-action cues) when attributed by close co-workers, as human "sensors," can be evaluated in a basic human trust relationship. Hosmer (1995) characterized trust as "the expectation by one person, group or firm of ethically justifiable behavior—that is, morally correct decisions and actions based upon ethical principles of analysis" (p. 399). However, trust can be violated, and a trustee can renege on obligations. When competence-based trust is violated, a trustor may feel a breach of psychological contract against the collective goal (Piccoli & Ives, 2003). When integrity-based trust is violated, it may trigger a natural attribution process involving extensive cognitive and affective reactions among trustors toward the trustee (Ho *et al.*, in press).

Tyler and Degoey (1996) proposed trustworthiness attributions of subordinates to their authority figures (e.g., supervisors and ad hoc team leaders). They found that in "instrumental models," subordinates (i.e., the trustors) tend to attribute trustworthiness based on the supervisor's *competence* (p. 343), while in "relational models," subordinates tend to draw on the implied *benevolence in the motives of authorities*. Tyler and Degoey (1996) specifically mentioned that "trustworthiness attribution" of authority is predominantly made on the basis of the *relational* aspects, suggesting that focal actors (e.g., the authority figures) will be trusted simply because of their role, even with little

prior knowledge of their competence. Both the duration and hierarchy of relationships can influence feelings of trust toward an authority figure, and organizations generally rely on the implied trustworthiness of employees. However, structure, duration, and hierarchy of relationships are not guarantees of trust. Just the opposite, the impact of the insider threat risk increases with key employees who have authorized access to critical information, but their violation of trust may go undetected. Several emerging studies (Ho & Benbasat, 2014; Ho & Warkentin, 2017) provide evidence that groups unknowingly[2] trigger a peer attribution process toward the perceived trustworthiness of their team leader that can reflect evidence of intention to betray.

Insider threat comes in different forms, including unintentional mistakes, intentional errors, or patriotic motive, but always involve some aspect of betrayal. When insiders are motivated to betray their organization, the underlying causes can involve both dispositional and situational factors. *Personal* or *dispositional (internal) factors* that contribute to the motivation for committing sabotage, fraud or theft generally include malicious intent to enact revenge, motivation for power, or monetary gain. *Situational (external) factors* include aspects of the surrounding environment, e.g., one's learned skills or actions being evaluated (Lord & Smith, 1983), and one's social network or cultural influence from community or society.

Accordingly, we propose a contemporary perspective to evaluate trustworthiness by adapting the dyadic attribution mechanism (Ho & Benbasat, 2014) in group

---

[2] Unconsciously; not making a conscious choice such as reporting an incident or filling out a survey.

communication (Ho *et al.*, in press). The dyadic attribution model posits that, (a) individuals signal their communicative intent and information behavior in words and actions, (b) members of close groups make attributions associated with trustworthiness of focal actors based on communication cues embedded in group exchange, and (c) these observations can lead to correct attributions of downward shifts in perceived trustworthiness when focal actors initiate in actual "betrayal activities" (Levine & McCornack, 1991). *Trustworthiness* refers to the reliability, assessed quality or characteristic of the trustee (Meyerson *et al.*, 2006), and suggests that the output of that trustee is, and will remain, dependable, true, accurate, truthful, and uncompromised. Perceived trustworthiness, on the other hand, refers to the perception of a trustee, defined as a *generalized expectation concerning the degree of congruence between an actor's communicated intentions and behavioral outcomes, which remain reliable, ethical and consistent, and any fluctuation between perceived intentions and actions does not exceed the observer's expectations over time* (Ho & Benbasat, 2014; Hosmer, 1995). This definition highlights two important observational constructs of an attribution process: *consistency (of words with actions)* and *distinctiveness (of behavior)* when compared with baseline expectations.

Rather than examining and analyzing a perpetrator who has already violated trust, this perspective suggests taking a step back to examine an actor's perceived trustworthiness as attributed within his/her group and reflected in the group interaction. This theoretical stance focuses on identifying: 1) observed behavior of a focal actor; 2) the cause of that behavior, as attributed by his/her close work group based on observed

and exchanged communications; and 3) any resulting shift in the perceived trustworthiness—up or down—attributed by the group to some cause.

The following propositions take the lens of multi-level variance considerations of both individual level (a focal actor), and group level (observers) in organizational or social contexts (Burton-Jones & Gallivan, 2007). Social actors, as observers in close relationships, communicate and construct *meaning* within an organization (Bigley & Pearce, 1998). Through communication, peers in group interactions will naturally evaluate the trustworthiness of the focal actor (Hardin, 1996). The accuracy of the group's judgment depends on the level of group sensitivity to behavioral shifts and changes (Winship & Hardy, 1999). More precisely, perceptions of a focal actor's behavior occur at the individual level, but the unit of analysis is group members' interaction over time.

Propositions

The process of trustworthiness attribution *by an observer group* occurs in two stages: first, building a generalized expectation of observed behavior of a focal actor against which to gauge anomalies, and second, making attributions as to whether these anomalies impact perceived trustworthiness, specifically referring to the *integrity* of any focal actor. Attribution theory argues that observers observe a focal actor's behavior over time, making inferences about causes behind changes in behavior that signify something either *consistent*, *different* or *abnormal* (Heider, 1944; Kelley, 1973), and make attributions of the focal actor's trustworthiness accordingly. A perception of lowered (or sustained) trustworthiness is usually based on attributing anomalies to internal or external

causes. In group communication, when any actor engages in betrayal activities (i.e., observables of actual trustworthiness), this betrayal intent may be reflected and result in subtle changes of physical and informational communication behaviors (Ho & Benbasat, 2014).

> Proposition 1 (P1): *Insiders who engage in betrayal activities against their organization may unknowingly send signals in communication artifacts reflecting reduced trustworthiness.*

Trustworthiness and Betrayal

When trust is violated, the trustee (i.e., the focal actor) may no longer be perceived as trustworthy, and could have an sense of ethical dilemma due to betrayal or trust violation against the trustors (i.e., observers) (Ho *et al.*, in press). Betrayal is an act of trust violation by the trustee who has knowingly departed from the norms—assumed to govern the relationship with the trustor—and, thus causing harm to the trustor (Finkel *et al.*, 2002). Moreover, betrayal implies an act of deliberate disloyalty, which violates trust against the norms as perceived by the trustor in a close "implicit or explicit" relationship (Finkel *et al.*, 2002).

Our next set of the propositions establish that perceived trustworthiness can be attributed by the close work group of a focal actor, reflecting or indicating the status of actual trustworthiness or insider threat betrayal (Ho *et al.*, in press). More specifically, Mayer and Davis (1999); (1995) defined *trustworthiness* based on three specific factors: *benevolence*, *integrity* and *competence*; however, generally *competence* is an external or situational factor, with *integrity* being an internal or dispositional factor. Competence depicts a learned behavior subject to external elements. For instance, a person can

increase competence by acquiring a set of skills through training, or may fail because of situational factors outside one's control. On the other hand, integrity is internal, or dispositional. High integrity refers to a person behaving ethically, or choosing not to harm others even when the opportunity presents itself.

Ho and Benbasat (2014) found early evidence of stronger correlations between perceived trustworthiness and an actor's actual betrayal activities when anomalous behavior is attributed to internal causes and more closely associated with *integrity* as an internal attribute—rather than competence or benevolence. Ho *et al.* (in press) empirically tested that when an actor's ethical dilemma is manifested in an act of deception, such integrity-based trust violation can be unconsciously attributed by group members, and reflected in cognitive and affective group communication. In the context of *insider threats*, we exclude competence-based and benevolence-based trust as indicators for insider threat because an individual may be deemed incompetent and make unintentional, negligent errors, but can still be considered trustworthy. On the other hand, an individual who is deemed very competent could in fact behave unethically. Moreover, we propose that a person's benevolence is not a strong indicator for signaling insider threat because an individual with betrayal intent could still be "benevolent" to friends, co-workers or subordinates. As such, we argue that the single factor of perceived *integrity* (i.e., *integrity*-based trust) as a measure of perceived trustworthiness would be a more effective indicator in the context of insider threats.

We postulate that perceived *integrity*-based trust correlates strongly with indicators of betrayal (i.e., actual trustworthiness), providing direct implications of

insider threat against an organization; i.e., *benevolence* or *competence* does not involve intentional betrayal and is not a strong indicator of a trust violation. Someone who is perceived to have downward shifts in their trustworthiness, *ethically* speaking, may have already violated trust and engaged in betrayal activities. Thus, we postulate that:

> Proposition 2 (P2): *Benevolence, as a measure of trustworthiness, is not a strong indicator of an actor's violation of trust*.
>
> Proposition 3 (P3): *Competence, as a measure of trustworthiness, is not a strong indicator of an actor's violation of trust*.
>
> Proposition 4 (P4): *When a focal actor's perceived trustworthiness, in terms of integrity, is reduced, this may indicate an elevated tendency to betray the organization*.

Human Sensor's Attribution

Attribution refers to an act of ascribing a result to a cause. *Attribution theory* describes a process of people attributing (or assigning) causes of behaviors based on observed behavior (Heider, 1944). These attributions differ depending on different contexts, interpretations, and the actor being studied (Weiner, 1985). The attribution of observed behavior can also be influenced by observers' judgment of whether the observed acted intentionally (internal or dispositional causality) or unintentionally (external or situational causality) (Kelley, 1973). To continue the previous analogy, regardless of a trustee's (i.e., the focal actor's) actual motivation to betray, attribution theory argues that others observe the trustee's behavior over time, making inferences about causes behind changes in behavior that may signify something abnormal (Kelley, 1973). Ho and Benbasat (2014) extended attribution theory, seeking to answer the specific question of "how the attribution mechanism works to signal a violation of trust

based on the communicative language of social actors in a virtual organization," and investigated how social actors' trustworthiness *can be attributed collectively by group members* based on "language as symbolic action." They noted that "in online communications, people must rely on fairly simple rudimentary evidence (e.g., words and actions) to decide whether to trust another party" (Ho & Benbasat, 2014, p. 1). Consistent with earlier attribution theory, Ho and Benbasat (2014) found support for the argument that co-workers often serve as "smart sensors," who can make intelligent inference and dynamic trustworthiness attributions based on the interplay of words and actions. Schwarz (2015) also described such subjective experiences, or meta-level thoughts—that support inference and judgment in the reasoning process—as metacognition. In other words, based on simple rudimentary evidence (i.e., words and action), members of a group can decide whether their impression is "likely to be accurate," or whether the impression is "consistent" with their baseline knowledge about a focal individual. Humans are able to make these meta-level judgments regardless of whether others share the same impression, or whether the information on which their judgment "came from a reliable source" (p. 204). We thus postulate that:

> Proposition 5 (P5): *Humans as "smart sensors" can understand and interpret subtle communication signals associated with reduced trustworthiness*.

Moreover, a baseline expectation of normal behavior is a basic prerequisite for group's observation in close relationships. Only through a close relationship in a group with a focal actor can observers develop some baseline expectation and measurement against which to compare anomalous behavior (Rempel *et al.*, 1985). When an actor's

behavior changes, an actor's close associates will perceive and assign meaning to such behavioral changes. Thus, we further postulate that:

> Proposition 6 (P6): *Close relationships in work or social groups may allow for observation of subtle cues that may indicate changes in trustworthiness*.

Consistency of Words and Actions

Kelley (1973) and others (Heider, 1944; Ho & Benbasat, 2014; Martinko & Thomson, 1998; Weiner, 1985) have included three information types in their attribution models. These three constructs are relevant to understanding how humans attribute trustworthiness in an organizational or group context: behavioral *consistency* of an actor's words and actions*, distinctiveness* of an actor's behavioral cues across time*,* and a level of observers' *group consensus* about causes for observed behaviors. That is, the evaluation of whether an actor's words are consistent with associated actions will lead to expectations of congruence between words and actions as expressed in communication artifacts becomes an important indicator of trustworthiness. *Consistency* refers to observers concluding that a person's words and behaviors are consistent (reliable and expected) over repeated interactions, thus would not raise any red flags in observers' inference of trustworthiness.

People will be known in general for consistent words and actions, or the lack of it, i.e., inconsistency (e.g., not following through on a promise). These patterns of behavior set up expectations for how other's behavior is judged. As noted earlier, attributions, associated with the lack of benevolence, unkindness, or incompetence due to a stream of mistakes or errors, are not good indicators of betrayal activities. We thus propose that the *consistency* of an actor's words and actions as interpreted by observers can indicate

perceived trustworthiness, and raise the following propositions regarding (in)consistency of words and actions, as attributed to indicate lower trustworthiness:

> Proposition 7 (P7): *The inconsistency between an actor's words (communicative intent) and actions, when attributed to lack of benevolence, is not an indicator of lower trustworthiness*.
>
> Proposition 8 (P8): *The inconsistency between an actor's words (communicative intent) and actions, when attributed to incompetence, is not an indicator of lower trustworthiness*.
>
> Proposition 9 (P9): *The inconsistency between an actor's words (communicative intent) and actions, when attributed to lack of integrity, can be an indicator of lower trustworthiness*.

As the perceptions of lower trustworthiness can signal insider betrayal, proposition 10 thus serves as a confirmatory proposition based on the establishment of propositions 4 and 9, arguing that an observer's perception of inconsistency between the focal actor's words and action may serve as an indicator of potential trust violations.

> Proposition 10 (P10): *The inconsistency between an actor's words (communicative intent) and actions, when attributed to lack of integrity, may indicate that the actor has a tendency to betray.*

Distinctiveness of Behavior

The basic premise of an intrusion detection system (IDS) is that anomalous behaviors signal some change that should be flagged for further investigation (Debar *et al.*, 1999; Patel *et al.*, 2013). Likewise, several attribution studies have also noted that if a person's cognitive response to certain stimulus is different from her/his responses to other similar stimuli, then there is a *distinctive* reaction (Martinko & Thomson, 1998). The focus of this inference framework does not differentiate on the type of stimuli (e.g., money or revenge) that might motivate insider betrayal, or the type of actor's cognitive response to the stimuli. Rather, it is focused on the groups' (i.e., smart human sensors')

ability to learn and to differentiate a focal actor's *distinctive* behavior from *regular* behavior across time when a betraying stimulus is present. Instead of observing the betraying stimulus, we focus on signals as observed by others. Ho and Benbasat (2014) defined *distinctiveness* as the extent to which an actor is attributed to behave *distinctively different to a stimulus when compared to a generalized expectation of his/her profiled behavior* over time. We thus can refer to *distinctiveness*, or a distinctive behavior, as a distinguishable, notable quality of an individual being evaluated, when compared to *established baseline behavior in similar situations*, with a *distinctive* outcome during communication.

That is, if an actor's behavior is noticeably different from his/her usual behavior, and the changes in that behavior are attributed to *external* (or, situational) causes, the causes of that change would be viewed as being outside of his or her control, and this actor would not be held intentionally responsible for the act. Hence, the actor's integrity would not be considered compromised and his/her perceived trustworthiness would not be adversely affected. We similarly argue that if a certain behavior is observed to be different from a focal actor's usual behavior, and the changes to that behavior are attributed to *internal* causes (i.e., due to a focal actor's choice or disposition), the actor could be held intentionally responsible for the act, hence his/her trustworthiness (i.e., *integrity*-based trust) would be attributed lower. Due to the earlier propositions that both benevolence and competence are not indicators of betrayal, we thus postulate only one proposition illustrating the relationships between the constructs of distinctiveness in actions and trustworthiness specifically referring to *integrity*-based trust violation.

Proposition 11 (P11): *When an actor acts noticeably different as compared to observers' baseline knowledge of the actor in similar situations, and the reason for such distinctive behavior is attributed to the actor (internal causality), the actor is likely attributed to have lower trustworthiness. By contrast, when an actor's behavior does not deviate from observers' baseline knowledge of the actor in other similar situations, and if such a behavior is attributed to be outside of the actor's control (external causality), the actor is likely attributed* ***not*** *to have lower trustworthiness*.

Group Consensus

Although it is unlikely that every actor in an organization would respond to a typical betrayal stimulus in the same way—with the same communicative cues, we argue that *group consensus* reflects a group's metacognitive inference process (Schwarz, 2015) in attributing cause of an observed behavior. This collective metacognitive inference process refers to people's ability to draw inference from each other's memory (Smith & Schwarz, 2016) and collectively know "that they know (collectively) something even though they cannot retrieve it at the moment… people's confidence in the accuracy of their knowledge and memories is indicative of actual accuracy (p. 207)." Ho and Benbasat (2014) noted this collective attribution or sense-making results a level of *group consensus*, which is the *extent to which group members are in agreement (or not) about a focal actor's observed behavior* under typical circumstances over time. Group consensus (or lack of) is a collective response that can be observed in the trustworthiness attribution of a focal actor.

Individual observers form their own baseline knowledge about an actor based on personal social interactions (Lewicki & Bunker, 1996). When an actor's words are consistent with his/her action, and no distinctive cues of unusual behavior are noted based on baseline expectations for similar situations, group members tend to reach agreement

about an actor's behaviors. In the next set of propositions, we observe the variation of group consensus in relation to the attributed cause of an actor's behavior, e.g., when *distinctiveness* in behavior occurs (based on established patterns of the actor's prior behavior), or when *inconsistency* is found between an actor's words and actions. Either or both of these may challenge observational agreement within groups, as baseline knowledge or "observational templates" about a focal actor's usual behavior may vary across multiple observers (Shaw, 1961; Steiner, 1964; Tuckman, 1965). Ho *et al*.'s (in press) experiments found evidence in the collective ability of groups to reflect and signal a deceptive actor in computer-mediated group communication. That is, groups will display more cognitive load, affective process, and use more words of negation in group communication when a deceptive actor is present. By comparing a deceptive actor's words with the words used by an interacting group, Ho *et al.* (in press) further identified that a deceptive actor will use more words of negation than other interacting group members in synchronous communication.

Lack of group consensus can also be an indicator that a focal actor's behavior has been collectively attributed to an internal (dispositional) causality, resulting in lower trustworthiness (i.e., as measured by *integrity*-based trust), which signals a higher risk of insider threat behavior. Likewise, when group agreement is observed over multiple interactions, the focal actor's behavior is more likely to have been attributed to external (situational) causality (e.g., the actor's competence), which would not result in lower trustworthiness (i.e., as measured by integrity-based trust). Thus, we postulate that group consensus—or lack thereof—is a collective response that can be observed to infer the

cause of an actor's distinctive behavior—being dispositional or situational—as associated with shifts in trustworthiness specifically referring to integrity.

> Proposition 12 (P12): *When groups do not reach consensus about a focal actor's distinctive behavior, and if the behavior is attributed to disposition, then the actor is likely perceived as less trustworthy*.

Moreover, lack of group consensus over multiple interactions is a phenomenon that can be observed to infer the actor's inconsistency between words and action as associated with perceived untrustworthiness. Thus, we postulate that,

> Proposition 13 (P13): *When groups do not reach consensus about a focal actor's inconsistency between words and actions, and if the behavior is likely attributed to disposition, then the actor is likely perceived as less trustworthy*.

Group Sensitivity

"*Group sensitivity*" affects a group's collective ability to sense and react to anomalous behavioral cues. While a group's general consensus on cause (i.e., dispositional vs. situational) based on observed behavior may be measurable, group dynamics can be quite unique and any group may be more or less sensitive to behavioral consistency or distinctive behaviors (Shaw, 1961; Steiner, 1964). *Group sensitivity* also represents the group's *degree of general awareness* toward the focal actor's communicative cues, e.g., consistency between words and actions and also any distinctive behavioral cues (Winship & Hardy, 1999). In the simplest terms, the degree of group sensitivity is bidirectional (e.g., suspiciousness vs. trustfulness), which could influence each member's attribution of trustworthiness in an exchanged message. Different observations by individual members will vary within a group. If even only one team member (an observer) has higher sensitivity to cues from a focal actor, this observer

can influence the group's collective view by sensitizing it to cues that might have been ignored.

Group sensitivity can be influenced by different degree of trustfulness across individual observers, or trustors. *Trustfulness* refers to a characteristic in the trustor. Every trustor may have his or her own level of trustfulness, and thus different characteristics of trustors would influence group behavior, attribution as well as agreement about observations and the cause of what is observed (Tuckman, 1965). *Trustfulness* is an inherent tendency to trust, which may cause *truth bias* (Street & Masip, 2015). The illusory *halo error*, or *halo effect* can be embedded in the observer (or, evaluator) (Cooper, 1981). That is, based on different degrees of *trustfulness*, an observer can be biased when being asked to make judgments, suggesting that *trustfulness* may influence the attribution process (Levine & McCornack, 1991; Tyler & Degoey, 1996). For example, a naïve observer tends to be biased toward believing a speaker is telling the truth (Bond & DePaulo, 2006; Street & Masip, 2015) due to heuristic processing of information (Kahneman & Tversky, 1972; Street & Masip, 2015). Every individual observer may have different degrees of *trustfulness* that influences his/her belief/opinion system. Different characteristics and *trustfulness* of individual observers may further influence group behavior, dynamics and agreement about observations. The data analyzed in Ho and Benbasat's (2014) study illustrates that when participants are asked to make an evaluation, their different levels of *trustfulness* might affect the group's overall *sensitivity*.

In an organizational context, employees may be more sensitized if they have experienced insider threats. Similarly, employees involved in IT Security may be more aware of potential risks of insider negligence or threat, and become more sensitized to anomalous cues. Regardless, different degrees of observers' trustfulness would likely lead to individual-based truth bias, and might influence the group's dynamics. However, the results of Ho *et al*.'s (in press) experiments found strong evidence that when analyzing group communication (in contrast to using survey instruments), group *sensitivity* to anomalous communicative cues was identifiable with statistical significance. This implies that the impact of the *halo effect* was insignificant. Thus, we further extended the attribution studies by adding the construct of *group sensitivity* that measures *the degree to which a group can sense and accurately interpret anomalous behavior of a focal actor* depending on different contexts.

In short, collective reactions to a focal actor's behavioral cues can be influenced by communication structure among observers, as well as an actor's leadership style, power structure, and power dynamics. As degrees of trustfulness—based on naïve or a suspicious nature—can influence how people determine the accuracy of their belief, we argue that "sensitivity" must be considered when evaluating a group's collective response and metacognition (Schwarz, 2015; Smith & Schwarz, 2016) to communication cues.

> Proposition 14 (P14): *The higher a group's sensitivity to an actor's behavioral changes, the more likely the group will detect inconsistencies between a focal actor's words and actions.*

Unfortunately, when one observer in a group has higher or lower sensitivity to behavioral cues, this can influence and reduce the likelihood of a group reaching

agreement about the focal actor's behavior, behavioral change, or its meaning, and can result in lower overall attribution of trustworthiness. We thus argue that even when group sensitivity is influenced by one or more group member(s), the aggregated *group sensitivity* will be tuned to notice anomalous cues of the focal actor. We further postulate that *group sensitivity* to the focal actor's behavioral cues is inversely related to group consensus in attributions of cause. Thus,

> Proposition 15 (P15): *The higher a group's sensitivity is to an actor's behavioral changes, the more likely this group will exhibit difficulty in reaching agreement on observed behaviors*.

A Research Model at Work

The above propositions aim at identifying patterns of insider betrayal based on the communication cues and artifacts exchanged between group members in either online or hybrid (face-to-face and online) settings. These 15 propositions are formulated based on a focal actor's perceived trustworthiness when an actor has a tendency to betray, or has already done so. Our framework requires two premises in the context of integrity-based trust violation. As the main dependent construct is perceived trustworthiness, as measured by integrity, this model first states that perceived trustworthiness, attributed collectively, can be indicative of actual trustworthiness (i.e., *betrayal*). This was empirically tested in Ho *et al.* (in press) experiments that indicate that perceived trustworthiness—as attributed by a close work group—is indicative of actual trustworthiness. Ho *et al.* (in press) empirically supported this main proposition with data collected between 2008 and 2015—that a focal actor's betrayal concealed in the deceptive activities were identifiable through groups' collective cognitive and affective

communication with statistical significance. Second, this model requires a panorama of the dynamics of complex interactions between observers and any actor in close relationship (Ho & Warkentin, 2017). The close relationships and interaction allows an opportunity for observers to assess and attribute an actor's perceived trustworthiness with observational accuracy. The panoramic (or, multi-faceted) view of the work group environment provides the research platform and premise for a more efficient approach to analyzing relationships among various constructs as proposed in the propositions and based on available communication cues and artifacts. This panoramic view of insider activities enables non-invasive routine monitoring, with no violation of employee privacy.

Discussion and Directions for Future Research

Insider threats represent a significant organizational problem where trusted insiders betray their colleagues and violate the collective interests of organizations. Current insider threat research often focuses on after-the-fact violations of security, access controls, or standard operating procedures. Due to privacy and legal concerns, as well as employees' fears of reprisal or resistance to whistleblowing, research on the role of group attribution in insider threat situations is lacking. Research on group attribution will move us beyond reliance on individual reports (whistleblowing) into understanding patterns of collective communication behavior. This paper enlightens the intelligence community that wishes to detect early warning signs of insider betrayal by building on sociotechnical works with both technical and behavioral evidence, tapping into the root cause of downward shifts in trustworthiness as assessed by interacting group members.

This inquiry extends Ho and Benbasat's (2014) dyadic attribution model, arguing that use of communication cues and artifacts to analyze attributed shifts in trustworthiness may suggest insider threat activities. This perspective further emphasizes the importance of *integrity*-based trust as a stronger predictor of downward shifts in trustworthiness when insider betrayal is involved. More research is encouraged to distinguish the efficacy of trustworthiness in moral and *ethical* situations, as well as the role of internal and external causality in this context. With limited communicated cues embedded in group exchanges signaling downward shift of perceived trustworthiness, more evidence is required to correlate betrayal activities with low trustworthiness. To advance on further research of insider threat detection, these propositions should be tested using communication artifacts and language-action cues from insider threat simulations.

Human perceptions are not fully reliable, due to the fact that not all information is available to the observers. However, human observation naturally leads to attribution of people's trustworthiness based on limited information and interactions. Our framework advances the conventional attribution model by refocusing attribution of an actor's trustworthiness when based on observed behaviors from "human sensors" within a group context. Moreover, this framework contributes theoretical arguments and provides analytical insights in the arena of insider threat detection. By understanding how "group sensitivity" and "group consensus" work and fluctuate in detecting insider threats, we may be able to derive a more timely aggregated assessment. Similarly, most attributions are *context*-specific, time-dependent, and combined with assessment regarding an actor's capability to maintain responsibility and accountability for achieving external goals. This

paper suggests new attention be paid to "human sensors" as an important element in future "design" of sociotechnical artificial intelligence computational systems. Future research can be expanded to the design and development of computational systems based on these theoretical stances.

References

Bigley, G. A., & Pearce, J. L. (1998). Straining for shared meaning in organizational science: Problems of trust and distrust. ***Academy of Management Review,*** *23*(3), 405-421.

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. ***Personality and Social Psychology Review,*** *10*(3), 214-234.

Cooper, W. H. (1981). Ubiquitous halo. ***Psychological Bulletin,*** *90*(2), 218-244.

Debar, e., Dacier, M., & Wespi, A. (1999). Towards a taxonomy of intrusion-detection systems. ***Computer Networks,*** *31*(8), 805-822.

FBI Press Release. (2001). FBI history famous cases: Robert Philip Hanssen espionage case. In Office, F. N. P. (Ed.): Federal Bureau of Investigation.

Finkel, E. J., Rusbult, C. E., Kumashiro, M., & Hannon, P. A. (2002). Dealing with betrayal in close relationships: Does commitment promote forgiveness? ***Journal of Personality and Social Psychology,*** *82*(6), 956-974.

Hardin, R. (1996). Trustworthiness. ***Ethics,*** *107*(1), 26-42.

Heider, F. (1944). Social perception and phenomenal causality. ***Psychological Review,*** *51*, 358-374.

Ho, S. M., & Benbasat, I. (2014). Dyadic attribution model: A mechanism to assess trustworthiness in virtual organization. ***Journal of the Association for Information Science and Technology,*** *65*(8), 1555-1576.

Ho, S. M., Hancock, J. T., & Booth, C. (in press). Ethical dilemma: Deception dynamics in computer-medicated group communication. ***Journal of the American Society for Information Science and Technology***.

Ho, S. M., Hancock, J. T., Booth, C., & Liu, X. (2016). Computer-mediated deception: Revealed by language-action cues in spontaneous communication. ***Journal of Management Information Systems,*** *33*(2), 393-420.

Ho, S. M., & Warkentin, M. (2017). Leader's dilemma game: An experimental design for cyber insider threat research. ***Information Systems Frontiers,*** *19*(2), 377-396.

Hosmer, L. T. (1995). Trust: The connecting link between organizational theory and philosophical ethics. ***Academy of Management Review,*** *20*(2), 379-403.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. ***Cognitive Psychology,*** *3*, 430-454.

Kelley, H. H. (1973). The process of causal attribution. ***American Psychology,*** *28*(2), 107-128.

Levine, T. R., & McCornack, S. A. (1991). The dark side of trust: Conceptualizing and measuring types of communicataive suspicion. ***Communication Quarterly**, 39*(4), 325-340.

Lewicki, R. J., & Bunker, B. B. (1996). Developing and maintaining trust in work relationships. In Kramer, R. M. & T. R. Tyler (Eds.), *Trust in Organizations: Frontiers of Theory and Research* pp. 114-139. Thousand Oaks, CA: Sage.

Liang, N., Biros, D. P., & Luse, A. (2016). An empirical validation of malicious insider characteristics. ***Journal of Management Information Systems**, 33*(2), 361-392.

Lord, R. G., & Smith, J. E. (1983). Theoretical, information processing, and situational factors affecting attribution theory models of organizational behavior. ***Academy of Management Review**, 8*(1), 50-60.

Martinko, M. J., & Thomson, N. F. (1998). A synthesis and extension of the Weiner and Kelley attribution models. ***Basic and Applied Social Psychology**, 20*(4), 271-284.

Mayer, R. C., & Davis, J. H. (1999). The effect of the performance appraisal system on trust for management: A field quasi-experiment. ***Journal of Applied Psychology**, 84*(1), 123-136.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. ***Academy of Management Review**, 20*(3), 709-734.

Meyerson, D., Weick, K. E., & Kramer, R. M. (2006). Swift trust and temporary groups. In Kramer, R. M. (Ed.), *Organizational Trust: A Reader* pp. 415-444. New York, NY: Oxford University Press.

Mitnick, K. D., & Simon, W. L. (2002). *The art of deception: controlling the human element of security*. Indianapolis, Indiana: Wiley.

Moore, A., Cappelli, D., Caron, T., Shaw, E., & Trzeciak, R. (2009). Insider theft of intellectual propertt for business advantage: A preliminary model: CERT Program and Software Engineering Institute.

Office of the Director of National Intelligence. (2014). *Insider threat*. The Office of the National Counterintelligence Executive, Office of the Director of National Intelligence

Patel, A., Taghavi, M., Bakytiyari, K., & Junior, J. C. (2013). An intrusion detection and prevention system in cloud computing: A systematic review. ***Journal of Network and Computer Applications**, 36*(1), 25-41.

Piccoli, G., & Ives, B. (2003). Trust and the unintended effects of behavioral control in virtual teams. ***Management Information Systems Quarterly**, 27*(3), 365-395.

Rempel, J. K., Holmes, J. G., & Zanba, M. P. (1985). Trust in close relationships. ***Journal of Personality and Social Psychology**, 49*(1), 95-112.

Schwarz, N. (2015). Metacognition. In Mikulincer, M., P. R. Shaver, E. Borgida, & J. A. Bargh (Eds.), *APA Handbook of Personality and Social Psychology* Vol. 1: Attitudes and Social Cognition, pp. 203-229. Washington, D.C.: American Psychological Association.

Shaw, M. E. (1961). Group dynamics. ***Annual Review of Psychology**, 12*, 129-156.

Smith, R. W., & Schwarz, N. (2016). Metacognitive inferences from other people's memory performance. ***Journal of Experimental Psychology: Applied**, 22*(3), 285-294.

Steiner, I. D. (1964). Group dynamics. ***Annual Review of Psychology**, 15*, 421-446.
Street, C. N. H., & Masip, J. (2015). The source of the truth bias: Heuristic processing? ***Scandinavian Journal of Psychology**, 56*, 254-263.
Tate, J. (2013, August 21, 2013). Judge sentences Bradley Manning to 35 years, Tate, J., *WashingtonPost*.
Times, T. N. Y. (2014, January 1, 2014). Edward Snowden, Whistle-Blower. *The New York Times*.
Toxen, B. (2014). The NSA and Snowden: Securing the all-seeing eye. ***Communication of the ACM**, 57*(5), 44-51.
Tuckman, B. W. (1965). Developmental sequence in small groups. ***Psychological Bulletin**, 63*(6), 384-399.
Tyler, T. R., & Degoey, P. (1996). Trust in organizational authorities: The influence of motive attributions on willingness to accept decisions. In Kramer, R. M. & T. R. Tyler (Eds.), *Trust in Organizations: Frontiers of Theory and Research* pp. 331-356. Thousand Oaks, CA: Sage.
Verizon. (2016). 2016 Data breach investigation report (pp. 1-85): Verizon.
Verizon. (2017). 2017 Data breach investigation report (pp. 1-76): Verizon.
Wall, D. S. (2012). Enemies within: Redefining the insider threat in organizational security policy. ***Security Journal**, 26*(2), 107-124.
Weiner, B. (1985). An attributional theory of achievement motivation and emotion. ***Psychological Review**, 92*(4), 548-573.
Willison, R., & Warkentin, M. (2013). Beyond deterrence: An expanded view of employee computer abuse. ***MIS Quarterly**, 37*(1), 1-20.
Winship, G., & Hardy, S. (1999). Disentangling dynamics: group sensitivity and supervision. ***Journal of Psychiatric and Mental Health Nursing**, 6*(4), 307-312.